ACADEMIE DE MONTPELLIER

UNIVERSITE MONTPELLIER II

(Sciences et Techniques du Languedoc)

STATISTIQUE DES SCIENCES DE LA VIE ET DE LA SANTE

MASTER 2

Parcours professionnel

Méthodes Statistiques des Industries Agronomiques, Agroalimentaires et Pharmaceutiques

MEMOIRE sur le stage :

Sélection Génomique Multivariée chez le Palmier à Huile

- effectué du 01/03/14 au 31/08/14
- au CIRAD – La Recherche Agronomique pour le Développement
- sous la direction de : David Cros
- par : Alexandre H.D. Marchal
- soutenu le : 10/09/2014
- devant la commission d'examen : Jean-Noël Bacro
  Christelle Reynes
  Robert Sabatier

Résumé

La sélection génomique (Meuwissen *et al.*, 2001) s'impose aujourd'hui comme une méthode performante en amélioration des animaux d'élevage et des cultures. Le présent mémoire traite de l'utilisation de marqueurs SSR dans un programme de sélection récurrente réciproque du palmier à huile (*Elaeis guineensis* Jacq.), avec des données empiriques. Nous comparons différents modes de calculs de l'apparentement moléculaire à partir des données SSR afin d'optimiser le modèle G-BLUP. Nous montrons que le modèle G-BLUP basé sur la matrice d'apparentement d'Eding et Meuwissen (2001) est plus performant que le modèle BLUP basé sur le pédigrée, ou que le modèle G-BLUP basé sur une matrice d'apparentement de VanRaden (2007). Grâce à la variance d'erreur de prédiction (VEP), nous observons que le G-BLUP multivarié apporte une meilleure précision de sélection que le G-BLUP univarié pour prédire l'aptitude générale à la combinaison (AGC) des géniteurs lors d'un test génétique. Cependant la corrélation très forte entre les AGC prédites par les modèles univariés et multivariés est contradictoire avec l'amélioration élevée estimée par la formule de la VEP. Enfin, nous déterminons le nombre de marqueurs SSR minimal pour assurer les performances du modèle G-BLUP, mais aussi du modèle BayesB univarié. Nous observons que le modèle BayesB univarié fonctionne mieux à faible densité de marqueurs SSR que le modèle G-BLUP. Nous constatons également que la densité de marqueurs minimale à partir de laquelle les modèles génomiques surpassent le modèle basé sur le pédigrée dépend de la population (130 SSR pour le groupe A et 60 SSR pour le groupe B).

Mots-clefs : *Elaeis guineensis*, sélection récurrente réciproque, sélection génomique, SSR, G-BLUP, modèle mixte multivarié, données empiriques, aptitude générale à la combinaison.

Abstract

Genomic selection (GS) (Meuwissen *et al.*, 2001) is now considered as an efficient method for animal and plant breeding. This report deals with GS using SSR markers in a reciprocal recurrent selection program, for oil palm (*Elaeis guineensis* Jacq.) breeding. We used empirical data. Several ways to calculate molecular relationship using SSR data were compared, in order to optimize G-BLUP model. We showed that G-BLUP model based on Eding and Meuwissen (2001) relationship matrix was more efficient than BLUP model based on pedigree, or than G-BLUP model based on a VanRaden (2007) relationship matrix. We used the prediction error variance (PEV) to calculate accuracy of selection, and we observed that multivariate G-BLUP had a higher accuracy than univariate G-BLUP to predict general combining ability (GCA) of genitors evaluated in progeny tests. However, strong correlation between GCA predicted by univariate and multivariate models is inconsistent with high improvement estimated by PEV formula. Finally we determined the minimal number of SSR markers to ensure G-BLUP efficiency, and also univariate BayesB efficiency. We observed that univariate BayesB worked better at low SSR marker density than G-BLUP model. We also concluded that minimal genomic marker density from which genomic models outperformed pedigree-based model depended on the population (130 SSR markers for group A and 60 SSR markers for group B).

Keywords : *Elaeis guineensis*, reciprocal recurrent selection, genomic selection, SSR, G-BLUP, multivariate mixed model, empirical data, general combining ability.

Remerciements

Ce stage et la réalisation de ce mémoire n'auraient jamais été possibles sans la supervision de David Cros, chercheur en génétique quantitative au Cirad et améliorateur palmier à huile. Merci à lui pour sa patience et sa disponibilité à toutes les étapes de mon travail. Je tiens également à le remercier pour l'appui qu'il m'a apporté à une période cruciale de la maturation de mon projet professionnel.

Mes remerciements vont aussi à l'équipe de PalmElit. C'est grâce au partenariat avec PalmElit que le projet « Méthodes de SAM », dans lequel s'inscrit mon stage et mon mémoire, est rendu possible.

Merci à Andrés Legarra, chercheur en génétique quantitative à l'INRA, pour avoir adapté la matrice de VanRaden au cas multi-allélique. Un volet entier de ce mémoire s'appuie sur ce travail inédit.

Je remercie Jean-Marc Bouvet, chercheur en génétique quantitative au Cirad, pour ses commentaires qui ont aiguillé ma réflexion et m'ont permis de mieux comprendre la méthodologie en génétique quantitative.

Merci à Marie Denis, chercheur en statistique au Cirad, pour ses explications tant orales qu'écrites sur le modèle mixte et surtout sur le modèle Bayésien.

Enfin, je remercie toute l'équipe AGAP du Cirad qui a contribué à faire de ce stage une expérience aussi riche sur le plan professionnel qu'humain.

Table des matières

Table des illustrations

Glossaire

AGC : Aptitude générale à la combinaison.

AIC : *Akaike information criterion* – Critère d'information d'Akaïké.

AIS : *Alikeness in state* – Identité par état.

ANOVA : *Analysis of variance* – Analyse de variance.

ASC : Aptitude spécifique à la combinaison.

BIC : *Bayesian information criterion* – Critère d'information Bayésien.

BLUE : *Best linear unbiased estimation* – Meilleur estimateur linéaire non-biaisé.

BLUP : *Best linear unbiased prediction* – Meilleur prédicteur linéaire non-biaisé.

BV : *Breeding value* – Valeur génétique additive.

DL : Déséquilibre de liaison.

EBV : *Estimated BV* – BV estimée.

G-BLUP : Méthode BLUP utilisant la matrice d'apparentement G.

GEBV : *Genomic estimated BV* – Estimation génomique de la BV.

HSD : *Honest significant difference.*

IBD : *Identity by descent* – Identité par descendance.

MCMC : Markov Chain Monte-Carlo.

nearPD : fonction « *nearest positive definite* ».

QTL : *Quantitative trait locus.*

REML : *Restricted maximum likelihood* – Maximum de vraisemblance restreint.

SAM : Sélection assistée par marqueurs.

Sh : Gène codant pour l'épaisseur de la coque (pour *Shell*).

SG : Sélection génomique.

SNP : Marqueur « *Single-nucleotide polymorphism* ».

SRR : Sélection récurrente réciproque.

SSR : Marqueur « *Simple sequence repeats* » ou « microsatellite ».

TBV : *True breeding value* – BV réelle.

VC : Validation croisée.

VEP : Variance d'erreur de prédiction.

Notations mathématiques

a) Matrices

A : Matrice d'apparentement (*relationship matrix*) estimée à partir des données de pédigrée.

$\mathbf{A_{22}}$: Partie de A qui contient les données d'apparentement des géniteurs qui ont également été génotypés.

D : Matrice de dominance estimée à partir des données de pédigrée.

G : Matrice d'apparentement (*relationship matrix*) estimée à partir des données génotypiques :

- par la méthode de VanRaden (2007) étendue par Legarra (2014) : $\mathbf{G_{VR}}$;
- par la méthode de VanRaden (2007) normalisée par Forni *et al.* (2011) : $\mathbf{G_N}$;
- par la méthode d'Eding-Meuwissen (2001) : $\mathbf{G_{EM}}$.

H : Matrice contenant l'information sur l'apparentement généalogique et sur l'apparentement moléculaire.

I : Matrice identité.

R : Matrice de la variance-covariance des résidus d'un modèle linéaire mixte.

X : Matrice d'incidence (*design*) des effets fixes d'un modèle linéaire.

Z : Matrice d'incidence (*design*) des effets aléatoires, codée {0 (homozygote – allèle absent) ; 1 (hétérozygote) ; 2 (homozygote – allèle présent sur les deux chromosomes homologues)} pour les données génotypiques.

Γ : Matrice de la variance-covariance des effets aléatoires d'un modèle linéaire mixte.

b) Vecteurs

1 : Vecteur unitaire (1, 1, …,1).

b : Vecteur de paramètre des effets du modèle Bayésien.

e : Vecteur des résidus du modèle.

u : Vecteur de paramètres des effets aléatoires, prédits par BLUP.

y : Vecteur de la variable-réponse du modèle ($\mathbf{y_1}$, $\mathbf{y_2}$ si plusieurs variables-réponses).

β : Vecteur de paramètres des effets fixes, estimés par BLUE.

c) Scalaires

A : Effets additifs, ou *breeding-value* (BV).

a : Valeur génotypique additive.

AGC_x : Aptitude générale à la combinaison du parent x.

ASC_{xy} : Aptitude spécifique à la combinaison du croisement x×y.

C_{xy} : Covariance entre x et y.

D : Effets de dominance.

d : Valeur génotypique de dominance.

d_{xy} : Coefficient de fraternité entre les individus x et y.

df_θ : degrés de liberté (*degree of freedom*) de la loi du paramètre θ.

E : Déviance environnementale.

F_x : Coefficient de consanguinité (*inbreeding*) de l'individu x.

f_{xy} : Coefficient de parenté (*coancestry*) de Malécot entre les individus x et y.

G : Valeur génotypique.

h^2 : Héritabilité (au sens étroit).

I : Effets d'épistasie.

i : Intensité de sélection.

i_{ab} : variable indicatrice (0 ou 1) servant à calculer $S_{xy,l}$.

k_l : Nombre d'allèles au locus l.

L : Nombre total de loci de marqueurs moléculaires.

n : Nombre d'individus (*i.e.* de palmiers) dans l'échantillon.

P : Valeur phénotypique.

p ou p_k : Fréquence (de l'allèle k) dans la population.

q : Fréquence du second allèle dans la population.

R : Réponse à la sélection.

r : Précision (*accuracy*) de la prédiction.

S : Différentiel de sélection.

S_θ : Paramètre d'échelle (*scale*) de la t-distribution a priori du paramètre θ.

$S_{xy,l}$: Indice de similarité entre les individus x et y au locus l.

Y_{xy} : Valeur du croisement x × y.

α : effet de substitution du gène.

α_i : effet moyen de l'allèle A_i.

μ : Moyenne de la population.

σ^2 : Variance (σ_A^2 : variance additive, σ_e^2 : variance résiduelle, etc.).

d) Variables agronomiques

H : Vitesse de croissance du palmier en hauteur.

PM : Poids moyen d'un régime.

NR : Nombre de régimes.

PT : Poids total des régimes.

e) Lois de distributions

N : loi normale, loi normale multidimensionnelle.

Scaled-t : loi de Student (t-distribution) avec un paramètre d'échelle.

U : loi uniforme.

Bêta : loi bêta.

Gamma : loi gamma.

χ^{-2} : loi inverse-chi deux.

f) Ensembles

$\mathbb{R}$: Ensemble des réels.

Introduction

La production de semences et l'amélioration du palmier à huile sont des enjeux de taille puisque cette culture constitue la première ressource en huile alimentaire à l'échelle mondiale (indexmundi, 2014). Le Centre International de Recherche Agronomique pour le Développement (CIRAD), en partenariat avec l'entreprise semencière PalmElit, est engagé dans un programme de sélection de palmiers à huile.

L'apparition de la sélection génomique (Meuwissen *et al.,* 2001) a ouvert un nouveau champ d'action en amélioration végétale. David Cros, généticien au CIRAD et directeur du présent mémoire, prépare une thèse de doctorat sur la sélection génomique appliquée au palmier à huile. Le mémoire s'insère dans ce contexte. Nous nous intéressons particulièrement au modèle linéaire mixte génomique G-BLUP.

Nous nous proposons d'optimiser le modèle G-BLUP appliqué à des données empiriques de palmier à huile, génotypés avec des marqueurs microsatellites. Notre objectif principal sera de montrer si un modèle multivarié augmente la précision de la sélection génomique par rapport au modèle actuel univarié. Nous chercherons également à identifier l'estimateur d'apparentement génomique le plus performant. Enfin, nous nous interrogerons sur le nombre de marqueurs moléculaires suffisants pour faire fonctionner le modèle G-BLUP.

1. Contexte

1.1. Notions de génétique quantitative

Pour comprendre le contexte d'amélioration du palmier à huile, il convient de préciser quelques notions fondamentales de génétique quantitative.

1.1.1. Apparentement

On définit f_{xy} le coefficient de parenté (*Malécot's coefficient of coancestry* ou *kinship*) comme la probabilité qu'en tirant au hasard un allèle neutre et autosomal (*i.e.* non sexuel) chez l'individu x, celui-ci soit identique par descendance (*identic by descent,* IBD) à celui tiré au hasard, au même locus, chez un individu y (Toro *et al.*, 2011). L'identité par descendance des deux allèles signifie qu'ils ont été hérités d'un ancêtre commun. Cela revient à dire que f_{xy} est la proportion d'allèles IBD sur l'ensemble des deux génomes.

On définit F_x le coefficient de consanguinité (*inbreeding coefficient*) comme la probabilité que deux allèles de l'individu x, à un locus donné, soient IBD. On montre que (Toro *et al.*, 2011) :

$$f_{xx} = 0.5 \times (1 + F_x)$$

On définit l'apparentement (*additive genetic relationship*) entre deux individus comme deux fois leur coefficient de parenté (Falconer et Mackay, 1996). Cet apparentement peut être synthétisé dans une matrice carrée $\{2f_{xy}\}$, que l'on appellera **A** si elle est estimée à partir du pédigrée, et **G** si elle est estimée à partir de marqueurs moléculaires. En pratique, lorsque l'on estime l'apparentement par des marqueurs, on ne peut pas différencier l'identité par descendance de l'identité par état (*alikeness in state*, AIS) (Toro *et al.*, 2011).

Enfin, on définit d_{xy} le coefficient de fraternité comme la probabilité qu'à un locus donné, les deux allèles de l'individu x soient IBD aux deux allèles de l'individu y. On peut obtenir une approximation de d_{xy} par la formule (Ovaskainen *et al.*, 2008) :

$$d_{xy} \approx f_{fx\,fy} f_{mx\,my} + f_{fx\,my} f_{mx\,fy}$$

Où fx, mx, fy et my sont respectivement la mère et le père des individus x et y. On définit **D** = $\{d_{xy}\}$ comme étant la matrice de dominance.

1.1.2. Valeur génétique

On pose (Falconer et Mackay, 1996) :

$$P = G + E$$

Avec P la valeur phénotypique d'un individu (*e.g.* son rendement agronomique), G sa valeur génotypique, et E la déviation environnementale. L'espérance sur la population de E vaut 0, et l'espérance de G vaut l'espérance de P. Décomposons la valeur génotypique (Falconer et Mackay, 1996) :

$$G = A + D + I$$

Tableau 1 : détail des calculs de la BV (Falconer, et al., 1996).

Génotype	Fréquence	Valeur génotypique	Valeur additive (BV)
A_1A_1	p^2	a	$\alpha_1 + \alpha_1 =$ 2 q α
A_1A_2	2pq	d	$\alpha_1 + \alpha_2 =$ (q - p) α
A_2A_2	q^2	- a	$\alpha_2 + \alpha_2 =$ - 2 p α

Avec A les effets génétiques additifs ou *breeding value* (BV) de l'individu, D les interactions entre allèles à un même locus, ou effets de dominance, et I les interactions entre allèles à des loci distincts, ou effets d'épistasie.

On définit la BV d'un individu comme la somme des BV à chacun des locus contrôlant le caractère (*quantitative trait loci*, QTL) (Falconer et Mackay, 1996) :

$$A = \sum_{QTL} (\alpha_i + \alpha_j)$$

Les valeurs α_i et α_j sont appelées effet moyen des allèles A_i et A_j (respectivement) à un QTL donné, leur somme est la BV à ce locus. Pour illustrer ce concept, considérons le cas où 2 allèles A_1 et A_2 peuvent occuper un même locus (par convention, A_1 augmente la valeur par rapport à A_2). La BV du génotype à ce locus se calcule à partir des fréquences alléliques (p et q) et des valeurs génotypiques (a et d), comme présenté dans le Tableau 1. La valeur génotypique additive a correspond à la moitié de l'écart entre G_{A1A1} et G_{A2A2}. La valeur génotypique de dominance d est l'écart entre la valeur de l'hétérozygote G_{A1A2} et la valeur génétique intermédiaire entre les 2 homozygotes A_1A_1 et A_2A_2. La valeur α est appelée effet moyen de substitution du gène, on le calcule ainsi (Falconer et Mackay, 1996) :

$$\alpha = \alpha_1 - \alpha_2 = a + d(q - p)$$

Finalement, la BV d'un individu égale le double de l'écart à la moyenne de sa descendance issue d'un grand nombre de croisements aléatoires (double car seule la moitié de ses allèles est transmise à sa descendance). Ainsi, dans une population en équilibre de Hardy-Weinberg, l'espérance des BV égale 0 (Falconer et Mackay, 1996).

La BV est intéressante car les palmiers sélectionnés seront utilisés comme parents et que la BV renseigne sur les effets génétiques moyens qu'ils transmettront à leur descendance, utilisée pour la production d'huile (Piepho *et al.*, 2008).

1.1.3. Aptitude à la combinaison

Soit Y_{xy} la valeur du croisement x×y. On pose le modèle suivant (Gallais, 1990) :

$$Y_{xy} = \mu + AGC_x + AGC_y + ASC_{xy}$$

Avec AGC_x et AGC_y (aptitudes générales à la combinaison) correspondant aux effets génétiques additifs transmis par les parents x et y à leur descendance (*i.e.* $AGC = ½\ BV$), et ASC_{xy} (aptitude spécifique à la combinaison) l'interaction propre au croisement x × y (effet de dominance).

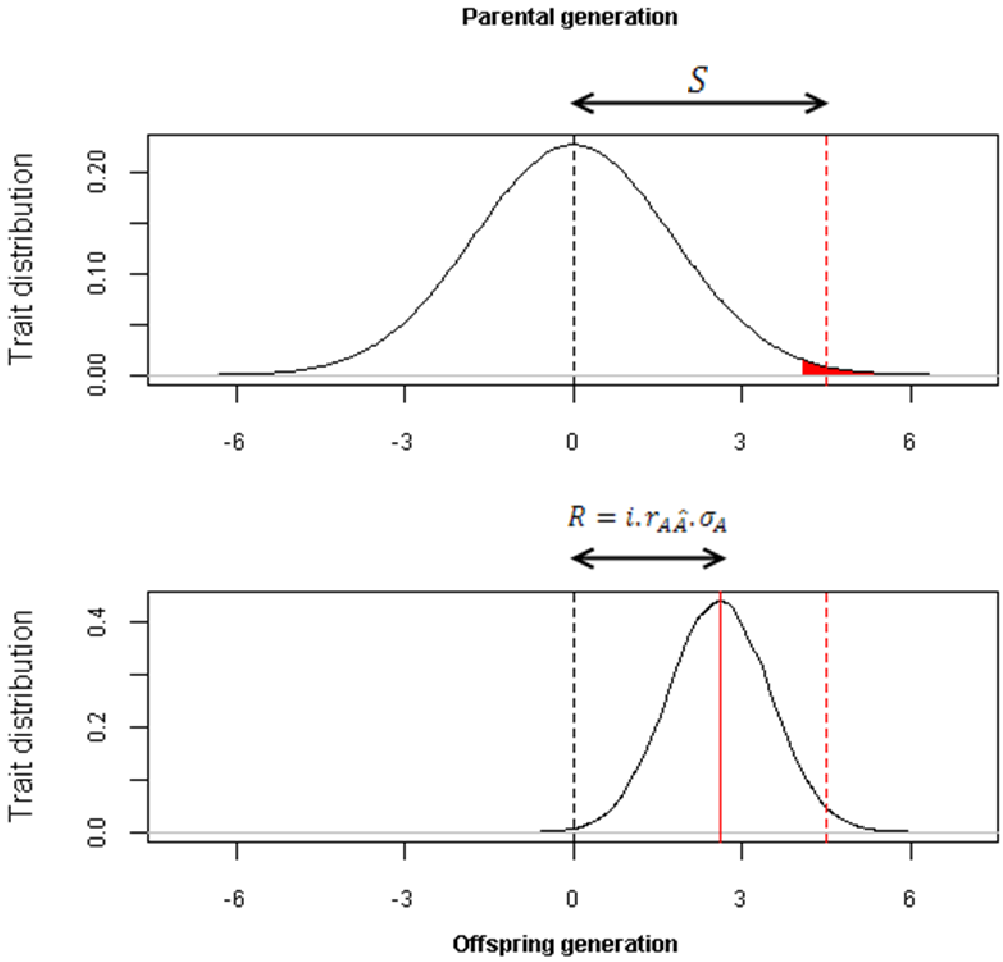


Figure 1 : Illustration de la réponse de sélection R et du différentiel de sélection S (Cros, 2013).

Les courbes gaussiennes représentent la distribution des valeurs prises par un caractère d'intérêt (e.g. un rendement agronomique) au sein de chacune des deux générations.

En haut : génération parentale. En pointillés noirs : la valeur moyenne du caractère pour l'ensemble des individus de la génération parentale. La partie en rouge correspond à la part des individus sélectionnés. En pointillés rouges : la valeur moyenne des individus sélectionnés.

En bas : génération de la descendance des parents sélectionnés. Le trait plein, en rouge, est la valeur moyenne du caractère pour les descendants.

1.1.4. Précision de sélection

Soit $\hat{A}$ le meilleur prédicteur linéaire de A, calculé en prenant en compte non seulement sa valeur phénotypique mais également celle des individus qui lui sont apparentés. On peut utiliser $\hat{A}$ comme un indice grâce auquel sélectionner les meilleurs parents à chaque génération.

Introduisons deux notions complémentaires. Le différentiel de sélection, noté S, est la différence entre la valeur phénotypique moyenne des parents sélectionnés et celle de la totalité de la génération parentale. La réponse à la sélection, notée R, est la différence entre la valeur phénotypique moyenne de la descendance des parents sélectionnés et celle de la totalité de la génération parentale. Ces deux notions sont illustrées par la Figure 1. La réponse à la sélection se calcule (Falconer et Mackay, 1996) :

$$R = i.r_{A\hat{A}}.\sigma_A$$

Avec i = S/σ_P, l'intensité de sélection, et σ_A^2 la variance d'additivité, issue de la décomposition :

$$\sigma_P^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2 + \sigma_E^2$$

Enfin, $r_{\hat{A}A}$ est le coefficient de corrélation de Pearson entre et $\hat{A}$. Par la suite, nous appellerons la *true breeding value* (TBV) et $\hat{A}$ l'*estimated* BV (EBV). Le coefficient $r_{TBV,EBV}$ est appelé précision de la sélection. Nous avons pour objectif de procéder à une sélection génomique. Dans le cas de la sélection génomique, l'estimateur $\hat{A}$ est appelé *genomic estimated BV* (GEBV) et $r_{A\hat{A}}$ devient $r_{TBV,GEBV}$. Par la suite, la précision sera simplement notée r. La précision peut être déduite de sa relation avec la variance d'erreur de prédiction (VEP) (voir 1.2.6. Erreur de prédiction).

Dans le cas particulier où l'EBV est estimée uniquement à partir du phénotype individuel (sélection massale), alors (Falconer et Mackay, 1996) :

$$r_{TBV,EBV} = \sqrt{h^2} = \frac{\sigma_A}{\sigma_P}$$

Le paramètre h² est appelé héritabilité au sens étroit (Gallais, 1990), ou simplement héritabilité.

Il est également possible de comparer la sélection génomique (GEBV) à une sélection opérée par une autre méthode (EBV) au moyen du coefficient de corrélation $r_{EBV,GEBV}$ appelé alors précision de prédiction (Hayes *et al.*, 2009; Lorrenz *et al.*, 2011).

Figure 2 : Couronne chargée de régimes d'un palmier à huile.

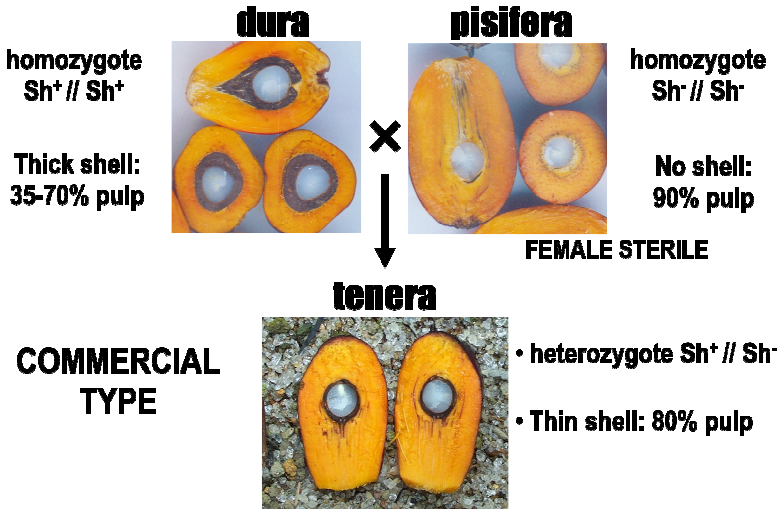


Figure 3 : Transmission mendélienne de l'épaisseur de la coque (Cros, 2013)

1.2. Le palmier à huile et son amélioration

1.2.1. Botanique et culture

Elaeis guineensis Jacq., le palmier à huile (Figure 2) est une plante monocotylédone, de la famille des *Arecaceae,* cultivée pour la production d'huile. Elle possède 16 paires de chromosomes homologues (2n=32) et la taille de son génome est estimée à 1743cM (Billotte *et al.*, 2005). Le palmier à huile est monoïque temporel, c'est-à-dire que les fleurs mâles et femelles se succèdent sur une même plante, favorisant son allogamie (*i.e.* fécondation croisée) (Durand-Gasselin *et al.*, 2000). La production de régimes, qui comportent 500 à 2000 drupes (*i.e.* fruits à noyau), est étalée tout au long de l'année. La plante commence à produire autour de sa 4ème année, et est en général exploitée pendant une vingtaine d'années, la hauteur de l'arbre devenant limitante. On extrait de la pulpe des fruits l'huile de palme rouge, et en moindre quantité on extrait de l'amande l'huile de palmiste. L'huile de palme est principalement valorisée en alimentation humaine (CIRAD, 2014). Les rendements en huile de palme, qui atteignent les 3.8 t/ha.an, supplantent ceux des autres oléagineuses cultivées (les rendements en huile du colza, tournesol, soja, etc., ne dépassent pas 0.7 t/ha.an) (MPOB, 2011). Cela a fortement contribué à ce que depuis 2003 la culture du palmier soit la première ressource en huile agroalimentaire à l'échelle mondiale. En 2014 sa production excède les 55 Mt (indexmundi, 2014). Le principal bassin de production se concentre en Asie du sud-est, surtout en Indonésie et en Malaisie. On distingue deux autres bassins secondaires : l'Amérique latine, notamment la Colombie, ainsi que l'Afrique tropicale, en particulier le Nigéria (indexmundi, 2014).

L'épaisseur de la coque est un caractère important car directement lié à l'épaisseur du mésocarpe (*i.e.* la pulpe) et donc à la teneur en huile des drupes du palmier. Ce caractère est notamment codé par un gène appelé Sh (pour *shell*) qui obéit aux lois de Mendel (Figure 3). Les drupes de l'homozygote Sh+ (dit *dura*) sont les moins riches en huile, les drupes de l'homozygotes Sh- (dit *pisifera*) sont les plus riches en huile. Malheureusement, les homozygotes Sh- sont abortifs et donc peu rentables en pratique. L'hétérozygote (dit *tenera*) présente des drupes assez riches en huiles, et n'est pas abortif. Les palmeraies à finalité de production d'huile sont donc plantées en *tenera*, obtenu par croisement *dura* × *pisifera* (Gascon et de Berchoux, 1964).

1.2.2. Déterminisme génétique des composantes du rendement

Le poids total (PT) de régimes produits par unité de temps (*e.g.* kg.arbre^{-1}.an^{-1}) est une composante importante du rendement du palmier à huile. Le PT est le produit de deux autres composantes, le nombre de régimes (NR) et le poids moyen des régimes (PM). Le NR et le PM sont négativement corrélés.

La sélection massale a été opérée au 20ème siècle de manière indépendante sur plusieurs sites : en Afrique (aire de répartition naturelle de la plante, population *dura*, *tenera* et *pisifera*), notamment en côte d'Ivoire et en République Démocratique du Congo, et en Asie (introduction en 1848 de 4 *dura*s). Elle a conduit à la divergence de plusieurs populations d'amélioration. Ces populations se répartissent en 2 groupes complémentaires.

Tableau 2 : Super-dominance du PT dans le cas d'un croisement A×B. Valeurs numériques choisies à titre d'exemple.

Groupe	NR (régimes/arbre.an)	PM (kg/régime)	PT (kg/arbre.an)
A	6	16.7	6×16.7 = 100
B	12	8.3	12×8.3 = 100
A×B	(6 + 12)/2 = 9	(16.7 + 8.3)/2 = 12.5	9×12.5 = **112.5**

Tableau 3 : Récapitulatif des caractéristiques des groupes A et B tels qu'ils sont actuellement utilisés dans l'amélioration génétique.

	groupe A	groupe B
Type	*dura*	*pisifera*
Rôle	♀	♂
NR	-	+
PM	+	-
Population	*e.g.* Deli	*e.g.* La Mé

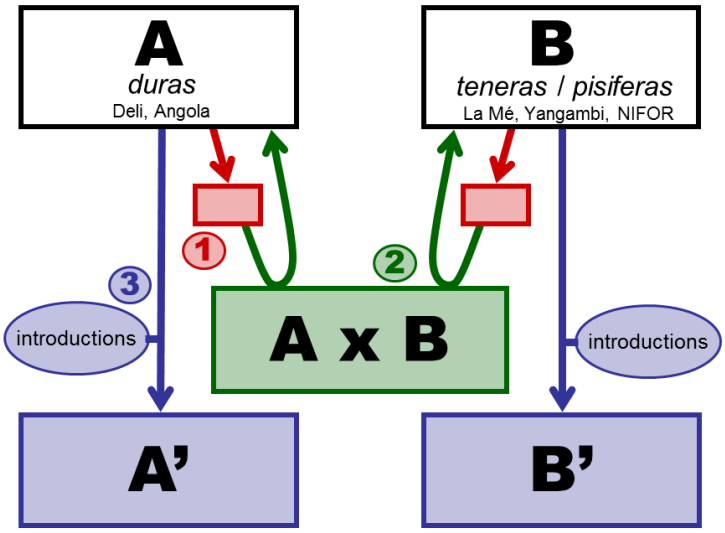


Figure 4 : Schéma de la SRR (Cros, 2013).

Le groupe A rassemble les populations qui produisent un petit nombre de gros régimes (Deli, Angola), et le groupe B rassemble celles qui produisent un grand nombre de petits régimes (*e.g.* La Mé, Yangambi) (Durand-Gasselin *et al.*, 2000). L' « Expérience internationale », plantée de 1950 à 1953, à l'initiative de l'Institut de Recherche pour les Huiles et Oléagineux (Gascon et de Berchoux, 1964), a permis de conclure quant à l'additivité de l'hérédité des caractères NR et PM, c'est-à-dire que la valeur génétique d'un descendant pour le caractère NR (respectivement PM) peut être prédit par moyenne des valeurs génétiques de ces parents pour le caractère NR (respectivement PM). Le PT est dit super-dominant dans le cas d'un croisement A × B. Cela signifie que la moyenne du PT de la descendance A × B dépasse la moyenne du PT de ses parents.

Le Tableau 2 illustre l'origine de cette super-dominance avec une application numérique. Les valeurs prises par les variables pour chacun des groupes sont cohérentes avec la réalité (Cros, 2013), mais ont été arrangées de façon à ce que les parents du groupe A et B aient exactement le même PT. Les NR et PM du croisement A × B sont simplement calculés par moyenne des NR et PM parentaux. Par multiplication, on obtient bien un PT pour la descendance supérieur à la moyenne des PT parentaux. On parle ainsi de complémentarité des caractères NR et PM. L'intérêt de l'additivité des caractères NR et PM est de permettre la prédiction du PT des croisements entre les parents testés en croisement et donc de pouvoir sélectionner les parents donnant les croisements avec le plus grand PT.

1.2.3. Production de semences

Les objectifs de la production de variétés commerciales sont :

- de croiser des *duras* avec des *pisiferas* pour ne produire que des *teneras* à forte teneur en huile ;
- de croiser le groupe A avec le groupe B pour maximiser le PT.

Ces caractères sont donc artificiellement associés : les *duras* du groupe A sont croisés avec les *pisiferas* du groupe B (Durand-Gasselin *et al.*, 2000). Les *duras* n'étant pas abortifs, ils sont naturellement utilisés comme porte-graines. Un récapitulatif des caractéristiques des deux lignées croisées pour la production de semences est proposé dans le Tableau 3.

1.2.4. Sélection récurrente réciproque

L'amélioration variétale poursuit plusieurs objectifs, le plus important étant d'augmenter le rendement agronomique. En moyenne pendant 40 ans, ce gain de rendement a dépassé les 1% par an. Parmi les autres enjeux de l'amélioration, soulignons la lutte contre la fusariose (*Fusarium oxysporuym* f.sp. *elaeidis*), la croissance en hauteur faible et la qualité de l'huile (Durand-Gasselin *et al.*, 2000).

L'amélioration variétale est pratiquée selon un schéma de sélection récurrente réciproque (SRR). Au sein des populations A et B, on présélectionne des géniteurs, sur les critères les plus héritables (croissance en hauteur faible, proportion d'huile dans la pulpe et épaisseur du mésocarpe) (Figure 4, étape 1). Ces géniteurs sont croisés, leur descendance hybride A × B est évaluée sur l'ensemble des critères (Figure 4, étape 2).

On sélectionne ainsi les meilleurs géniteurs, que l'on autoféconde et/ou que l'on croise pour produire une nouvelle génération des populations A et B (Figure 4, étape 3). Cette méthode présente certaines contraintes : (i) en termes de temps et d'espace, un cycle de sélection mobilise entre 1000 et 2000 ha pendant 25 ans (Durand-Gasselin *et al.*, 2000) ; et en conséquence (ii) seul un petit échantillon de géniteurs potentiel est testé (carrés rouges sur la Figure 4, prélevés dans les populations A et B).

1.2.5. Estimation de la BV : le modèle BLUP

Un modèle linéaire mixte est un modèle linéaire qui mélange des effets fixes et des effets aléatoires. Intéressons-nous à l'écriture générale du modèle linéaire mixte (Henderson, 1975) :

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}$$

Avec $\mathbf{y}$ le vecteur d'observation, $\mathbf{X}$ la matrice d'incidence des effets fixes, $\boldsymbol{\beta}$ le vecteur de paramètre des effets fixes, $\mathbf{Z}$ la matrice d'incidence des effets aléatoires, $\mathbf{u}$ et $\mathbf{e}$ des vecteurs d'effets aléatoires, tels que (Mrode, 2005) :

$$\mathbf{u} \sim \mathrm{N}(0, \boldsymbol{\Gamma}) \qquad \text{et} \qquad \mathbf{e} \sim \mathrm{N}(0, \mathbf{R})$$

Avec $\mathbf{u}$ et $\mathbf{e}$ indépendants. La méthode REML (*restricted maximum of likelihood*) est utilisée pour estimer les variances-covariances $\boldsymbol{\Gamma}$et $\mathbf{R}$ (Gilmour *et al.*, 2009). Puis les BLUE (*best linear unbiased estimations*) $\widehat{\boldsymbol{\beta}}$ et les BLUP (*best linear unbiased predictors*) $\widehat{\mathbf{u}}$ sont déduit de la résolution des équations du modèle mixte d'Henderson (Henderson, 1975) :

$$\begin{bmatrix} \mathbf{X}^{\mathrm{t}}\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^{\mathrm{t}}\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^{\mathrm{t}}\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^{\mathrm{t}}\mathbf{R}^{-1}\mathbf{Z} + \boldsymbol{\Gamma}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{\mathrm{t}}\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^{\mathrm{t}}\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

Le modèle dit BLUP est une application du modèle linéaire mixte. Il est utilisé pour prédire la valeur génétique additive ou BV des parents des individus observés. Ce modèle est utilisé pour analyser un essai dans lequel des palmiers des groupes A et B sont croisés, afin d'attribuer à chaque géniteur les performances de sa descendance et des autres palmiers qui lui sont apparentés. Le modèle BLUP consiste à poser (Mrode, 2005) :

$$\mathbf{R} = \mathbf{I}\sigma_{\mathrm{e}}^2$$

$$\boldsymbol{\Gamma} = \mathbf{A}\sigma_{\mathrm{A}}^2$$

Avec $\mathbf{I}$ la matrice identité, σ_{e}^2 la variance résiduelle, $\mathbf{A}$ la matrice d'apparentement au sein de la population parentale (Cf 1.1.1. Apparentement), et σ_{A}^2 la variance d'additivité (Cf 1.1.3. Sélection et précision) (Piepho *et al.*, 2008). Ainsi, il ne reste que σ_{e}^2 et σ_{A}^2 à estimer par REML (de los Campos *et al.*, 2013).

Les BV, considérées comme des effets aléatoires, sont donc prédites. Toutefois par abus de langage on parle souvent d'estimations (Walsh, 2013). Le vecteur $\widehat{\mathbf{u}}$ regroupe ainsi l'ensemble des EBV des individus impliqués dans le calcul de la matrice **A**, y compris les parents des individus observés.

1.2.6. Erreur de prédiction

La variance d'erreur de prédiction (VEP) peut être interprétée comme la part de variance additive qui n'est pas prise en compte par la prédiction du modèle BLUP (Mrode, 2005). Inversons (inverse généralisé) la matrice de coefficients de l'équation du modèle mixte :

$$\begin{bmatrix} \mathbf{X}^t\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z} + \mathbf{\Gamma}^{-1} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{C^{11}} & \mathbf{C^{12}} \\ \mathbf{C^{21}} & \mathbf{C^{22}} \end{bmatrix}$$

Si $\mathbf{C}_{xx}^{\mathbf{22}}$ est le x-ième terme de la diagonale de $\mathbf{C^{22}}$ alors (Mrode, 2005; Clark *et al.*, 2012) :

$$\mathrm{VEP}_x = \mathbf{C}_{xx}^{\mathbf{22}}\sigma_e^2$$

La VEP est reliée à la précision de sélection par la relation :

$$\mathrm{VEP}_x = (1 - r_x^2)\mathbf{\Gamma}_{xx}$$

Où r_x est la précision de sélection pour l'individu x. Si le BLUP est une BV, alors $\mathbf{\Gamma}_{xx} = 2f_{xx} \times \sigma_A^2 = (1+F_x)\,\sigma_A^2$ (Toro *et al.*, 2011). Ainsi (Gilmour *et al.*, 2009; Daetwyler *et al.*, 2013) :

$$r_x = \sqrt{1 - \frac{\mathrm{VEP}_x}{(1 + F_x)\sigma_A^2}}$$

1.2.7. Modèle mixte multivarié

Un modèle multivarié est un modèle qui explique simultanément plusieurs variables réponses. Il permet de valoriser la corrélation de ces variables pour gagner en précision (Jia et Jannink, 2012). Le modèle mixte multivarié a pour écriture matricielle (Mrode, 2005) :

$$\begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{bmatrix} = \begin{bmatrix} \mathbf{X_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{X_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{bmatrix} + \begin{bmatrix} \mathbf{Z_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{Z_2} \end{bmatrix} \begin{bmatrix} \mathbf{u_1} \\ \mathbf{u_2} \end{bmatrix} + \begin{bmatrix} \mathbf{e_1} \\ \mathbf{e_2} \end{bmatrix}$$

Avec (Mrode, 2005) :

$$\begin{bmatrix} \mathbf{u_1} \\ \mathbf{u_2} \end{bmatrix} \sim \mathrm{N}\left(0, \begin{bmatrix} \sigma_{A1}^2 & C_{A12} \\ C_{A12} & \sigma_{A2}^2 \end{bmatrix} \otimes \mathbf{A}\right)$$

Où $\otimes$ est le produit matriciel de Kronecker, σ_{A1}^2 et σ_{A2}^2 sont les variances génétiques additives des effets directs des caractères 1 et 2, et C_{A12} la covariance génétique additive entre ces caractères. De même, comme pour la MANOVA (Université de Toulouse, 2014):

$$\begin{bmatrix} \mathbf{e_1} \\ \mathbf{e_2} \end{bmatrix} \sim \mathrm{N}\left(0, \begin{bmatrix} \sigma_{e1}^2 & C_{e12} \\ C_{e12} & \sigma_{e2}^2 \end{bmatrix} \otimes \mathbf{I}\right)$$

Où σ_{e1}^2 et σ_{e2}^2 et C_{e12} forment la matrice de variance-covariance des résidus.

1.3. La sélection génomique

1.3.1. Principe et intérêt de la SG

La sélection assistée par marqueurs (SAM) a d'abord été utilisée pour suivre les QTL. Lorsqu'un marqueur est identifié comme étant en déséquilibre de liaison (DL) avec un QTL majeur, on peut suivre le gène d'intérêt grâce à ce marqueur. Cependant, lorsque le caractère ne s'explique pas par quelques gènes majeurs mais par un grand nombre de gènes apportant chacun une petite contribution, cette méthode montre ses limites (Lorrenz *et al.*, 2011). L'article de Meuwissen *et al.* (Meuwissen *et al.*, 2001) pose les fondements de la sélection génomique (SG). L'idée de ces auteurs est de valoriser l'information véhiculée par un grand nombre de marqueurs moléculaires, répartis sur l'ensemble du génome. La SG est donc une SAM (Nakaya et Isobe, 2012). Les QTL ne sont plus identifiés, on compte simplement sur le DL entre ceux-ci et les marqueurs. Les marqueurs doivent donc être suffisamment denses pour capter l'effet d'un maximum de QTL (Heffner *et al.*, 2009).

On appelle jeu d'apprentissage l'ensemble des individus pour lesquels on dispose à la fois d'un génotype et d'un phénotype, et l'on nomme candidats à la sélection l'ensemble des individus pour lesquels on ne dispose que du génotype. Le modèle de SG permet de prédire une GEBV aux candidats à la sélection simplement sur la base de leurs données moléculaires (Lorrenz *et al.*, 2011). De los Campos *et al.* (2013) concluent que, aussi bien avec des données simulées que des données empiriques, la SG est plus précise que la sélection traditionnelle basée sur le pédigrée.

En application au palmier à huile, on utiliserait comme jeux d'apprentissage les petits échantillons de géniteurs des populations A et B testés en croisement. Par la suite, on pourra estimer les GEBV d'un large jeu de géniteurs potentiels au sein des populations A et B en les génotypant (Cros *et al.* 2014).

Cela permettrait d'augmenter le différentiel de sélection, donc d'augmenter l'intensité de sélection. De plus, la SG pourrait être appliquée sur 2 voire 3 générations parentales successives, raccourcissant considérablement l'intervalle des cycles de sélection. En effet, les générations concernées passeraient d'un intervalle de plus de 20 ans (essai sur descendance) à un intervalle d'environ 6 à 8 ans (SG). Bien sûr la SG ne se substituera pas complètement aux essais sur descendance. A chaque essai sur descendance, le modèle s'enrichirait d'un nouveau jeu d'apprentissage, et devrait donc produire des prédictions de plus en plus précises (Heffner *et al.*, 2009).

1.3.2. Les modèles de SG

En génomique, il est courant de se retrouver confronté au problème « *large-p, small-n* », signifiant que l'on a beaucoup d'effets à évaluer et peu de degrés de libertés disponibles. La mise au point de méthodes de SG et leur comparaison a fait l'objet d'une abondante littérature ces dernières années, depuis l'article de Meuwissen *et al.* (2001).

Ces auteurs comparent une approche par *stepwise* et moindres carrés, deux méthodes Bayésiennes (nommées BayesA et BayesB), et un modèle mixte. Ce modèle mixte est alors utilisé pour prédire un effet associé à chaque allèle du marqueur.

Meuwissen *et al.* (2001) concluent quant à l'insuffisance de la méthode des moindres carrés, et obtiennent de meilleurs résultats avec BayesB et BayesA qu'avec le modèle mixte, sur des données simulées.

La revue de De los Campos *et al.* (2013) présente, entre autres, le modèle dit G-BLUP. Les auteurs constatent que le G-BLUP fonctionne bien dans la plupart des cas, cependant, ils soulignent la supériorité de BayesB, supériorité qui s'accroît quand on augmente le nombre de marqueurs (Meuwissen, *et al.,* 2009, Meuwissen et Goddard, 2010, *in* De los Campos *et al.* (2013)). Avec des données simulées, Jia et Jannink (2012) mettent en évidence l'importance de la structure génétique du caractère d'intérêt, les méthodes Bayésiennes (BayesA complètement hiérarchique, BayesCπ) s'avérant plus précises que G-BLUP pour un caractère déterminé par 20 QTL et G-BLUP s'avérant plus précis pour un caractère déterminé par 200 QTL. Dans leurs modèles univariés, cette différence s'accroit lorsque l'héritabilité du caractère augmente (de 0.1 à 0.5).

Avec des données réelles de palmier à huile, Cros *et al.* (2014) concluent à une absence d'effet significatif de la méthode de modélisation sur la précision des prédictions. Le test opéré est une ANOVA, chacune des répétitions étant une combinaison de validation croisée (VC), de groupe (A ou B), de variable agronomique mesurée. Cette conclusion les pousse à ne retenir que la méthode G-BLUP dans leur article. En effet cette méthode est moins contraignante en termes de temps et de puissance de calcul. Nous maintiendrons donc ce cap méthodologique en priorité.

1.3.3. G-BLUP et matrice d'apparentement génomique

Le modèle G-BLUP est similaire au modèle BLUP (Cf 1.2.5. Estimation de la BV : le modèle BLUP), sauf que l'on substitue la matrice **A** (calculée à partir du pédigrée) par une matrice **G** (estimée à partir des marqueurs moléculaires). Le vecteur $\hat{\mathbf{u}}$ regroupe ainsi l'ensemble des GEBV des parents. Remarquons que cette méthode ne donne pas d'estimation explicite des effets aux marqueurs mais directement une estimation de la valeur additive totale.

Notons que le choix de la matrice **G** et donc des estimateurs des coefficients de parenté repose sur des hypothèses fortes dérivant de la théorie de la génétique quantitative, hypothèses qui ne sont pas rencontrées en pratique (Piepho *et al.*, 2008). En particulier, dans les analyses génétiques, nous nous référons à une population de base ou population fondatrice composée d'individus que l'on considère non-apparentés (*i.e.* les individus qui la composent ne partagent pas de gènes IBD) (Forni *et al.*, 2011).

Intéressons-nous à différentes méthodes d'estimation de l'apparentement à partir de données de marqueurs moléculaires.

a) $\mathbf{G_{EM}}$ - Eding et Meuwissen

Soit $S_{xy,l}$ l'indice de similarité entre les individus x et y au locus l. Cet indice se calcule en utilisant les variables indicatrices i_{ab} telles que :

- Si l'allèle n°a du locus l de l'individu x est identique à l'allèle n°b du locus l de l'individu y, alors $i_{ab} = 1$
- Sinon, $i_{ab} = 0$.

Ainsi $S_{xy,l}$ se calcule (Eding et Meuwissen, 2001) :

$$S_{xy,l} = \frac{1}{4}[i_{11} + i_{12} + i_{21} + i_{22}]$$

La moyenne des $S_{xy,l}$ sur un grand nombre de loci est un estimateur du coefficient de parenté f_{xy}. S_{xy} est un estimateur sans biais à condition que les allèles fondateurs (*i.e.* les allèles de la population fondatrice) soient uniques (Eding et Meuwissen, 2001). Dans le cas contraire, il faudrait prendre en compte dans le calcul la probabilité que 2 allèles soient AIS, paramètre qui n'est pas accessible en pratique. On constate empiriquement que la matrice d'apparentement $\mathbf{G_{EM}}$ se calcule directement :

$$\mathbf{G_{EM}} = \frac{\mathbf{Z}\mathbf{Z}^t}{2 \times L}$$

Avec **Z** la matrice d'incidence génotypique, codée par le nombre de copies de l'allèle {0, 1, 2}, et L le nombre total de loci de marqueurs moléculaires. Soulignons que cette matrice est utilisée par Cros *et al.* (2014).

b) $\mathbf{G_{OF}}$ – VanRaden

Comme nous l'avons vu, la BV dépend de la structure génétique de la population de base (Cf 1.1.2. Valeur génétique et en particulier le Tableau 1). L'estimateur d'apparentement de VanRaden (2007) tient compte de cette nécessité en centrant la matrice d'incidence génotypique **Z** sur une matrice **P** qui contient le double des fréquences alléliques de la population de base :

$$\mathbf{P} = \begin{bmatrix} 2p_1 & \cdots & 2p_L \\ \vdots & \ddots & \vdots \\ 2p_1 & \cdots & 2p_L \end{bmatrix}$$

Cette méthode est utilisable dans le cas où 2 et seulement 2 allèles sont possibles pour chaque locus, par exemple dans le cas de marqueurs SNP (*single nucleotide polymorphism*). Les fréquences p_l correspondent aux fréquences de l'allèle le plus rare au locus l. Ces fréquences sont celles de la population que l'on considère être la population de base, ou population fondatrice. La méthode la plus simple, mais qui provoque des biais, est de les estimer en utilisant celles de l'échantillon. On définit la matrice d'apparentement $\mathbf{G_{OF}}$ (pour *observed frequencies*) la matrice produite en utilisant les fréquences alléliques de l'échantillon (VanRaden, 2007; Forni *et al.*, 2011):

$$\mathbf{G_{OF}} = \frac{(\mathbf{Z} - \mathbf{P})(\mathbf{Z} - \mathbf{P})^t}{2\sum_{l=1}^{L} p_l(1 - p_l)}$$

Legarra (2014) étend cet estimateur au cas où la population renferme plus de deux allèles différents pour un même locus, par exemple dans le cas de marqueurs *simple sequence repeats* (SSR) :

$$\mathbf{G_{OF}} = \frac{(\mathbf{Z}-\mathbf{P})(\mathbf{Z}-\mathbf{P})^{t}}{2\sum_{l=1}^{L}(1-\sum_{k=1}^{k_l} p_{lk}^{2})}$$

Avec l l'identifiant du locus (entre 1 et L), k l'identifiant de l'allèle (entre 1 et k_l au locus l).

c) $\mathbf{G_N}$ - VanRaden normalisé

Forni *et al.,* (2011) proposent une méthode pour « normaliser » la matrice d'apparentement de VanRaden (2008). Cette méthode vise à conduire à des estimations de la variance additive et de précision plus réalistes. On nomme cette nouvelle matrice d'apparentement $\mathbf{G_N}$ (pour *normalized*) :

$$\mathbf{G_N} = \frac{(\mathbf{Z}-\mathbf{P})(\mathbf{Z}-\mathbf{P})^{t}}{\{\mathrm{trace}[(\mathbf{Z}-\mathbf{P})(\mathbf{Z}-\mathbf{P})^{t}]\}/n}$$

Où n est le nombre d'individus (donc de lignes de la matrice $\mathbf{Z}$).

1.3.4. Le modèle G-BLUP multivarié

On peut étendre le modèle G-BLUP à l'analyse multivariée (Cf 1.2.7. Modèle mixte multivarié), l'intérêt étant, comme pour le modèle BLUP multivarié, de valoriser la corrélation génétique entre les variables pour augmenter la précision de sélection. Les modèles génomiques multivariés sont également capables de distinguer la corrélation génétique de la corrélation résiduelle (Gilmour *et al.*, 2009; Jia et Jannink, 2012).

Grâce à des données simulées, Jia et Jannink (2012) mettent en évidence l'importance de l'architecture génétique du caractère pour l'efficacité du modèle multivarié. Dans un modèle G-BLUP multivarié faisant intervenir un caractère peu héritable ($h^2 = 0.1$) et un caractère plus héritable ($h^2 = 0.5$), ils remarquent que seule la précision de sélection du caractère peu héritable augmente. Cette augmentation ne dépend pas du nombre de QTL contrôlant le caractère (20 ou 200). Par contre, plus la corrélation génétique entre les deux caractères est forte (de 0.1 à 0.9), et plus la précision obtenue pour le caractère peu héritable augmente.

Des résultats similaires sont obtenus par Calus et Veerkamp (2011), toujours en données simulées. La précision de sélection d'un modèle G-BLUP pour un caractère d'une héritabilité $h^2 = 0.9$ ne profite pas du passage au modèle multivarié, alors que la précision du modèle pour un caractère d'une héritabilité $h^2 = 0.6$ augmente. Ce gain de précision est d'autant plus élevé que la corrélation génétique avec le caractère très héritable est forte ($r = 0.25$, $r = 0.54$, $r = 0.75$).

Ces derniers auteurs mettent en garde contre une augmentation importante (+1379%) du temps de calcul pour le modèle G-BLUP lors du passage au modèle multivarié. De plus la demande en temps de calcul augmente quadratiquement avec le nombre d'individus du modèle.

Au contraire, ils remarquent que les modèles Bayésiens (BSSVS et BayesCπ) ne subissent qu'une légère augmentation du temps de calcul (+29%), et leur demande en temps de calcul augmente moins que linéairement avec le nombre d'individus (Calus et Veerkamp, 2011).

1.3.5. Le modèle BayesB

Les modèles BayesA et BayesB sont des modèles Bayésiens utilisés en sélection génomique. Ces modèles apportent une estimation de l'effet au marqueur. Commençons par présenter le modèle BayesA. Considérons le modèle suivant :

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zb} + \mathbf{e} \quad \text{avec} \quad \mathrm{e} \sim \mathrm{N}(0, \mathbf{I}\sigma_{\mathrm{e}}^2)$$

Avec **Z** la matrice d'incidence génotypique et **b** le vecteur des effets associés aux allèles. Définissons maintenant les distributions *a priori* de ce modèle. La variance résiduelle et l'origine suivent les lois suivantes (Rodriguez et de los Campos, 2013) :

$$\sigma_{\mathrm{e}}^2 \sim \chi^{-2}(\mathrm{S_e}, \mathrm{df_e})$$

$$\mu \sim \mathrm{U}(\mathbb{R})$$

Puis l'effet associé à l'allèle k suit une t-distribution (Rodriguez et de los Campos, 2013). La t-distribution permet de distinguer quelques marqueurs avec un grand effet d'un grand nombre de marqueurs avec un effet faible (Denis, 2012) :

$$\mathrm{b_k} \sim \text{Scaled-t}(\mathrm{df_b}, \mathrm{S_b})$$

On peut considérer la t-distribution comme un modèle à 2 niveaux (Denis, 2012):

$$\mathrm{b_k} \sim \mathrm{N}(0, \sigma_{\mathrm{b_k}}^2) \quad \text{avec} \quad \sigma_{\mathrm{b_k}}^2 \sim \chi^{-2}(\mathrm{S_b}, \mathrm{df_b})$$

On fixe par défaut $\mathrm{df_b} = 5$. Il reste à estimer $\mathrm{S_b}$ (Rodriguez et de los Campos, 2013) :

$$\mathrm{S_b} \sim \text{Gamma}$$

Enfin on passe à BayesB par l'introduction d'une probabilité π que le marqueur ait un effet non-nul. Ce nouveau paramètre doit être estimé (Rodriguez et de los Campos, 2013) :

$$\pi \sim \text{Bêta}$$

Cette approche théorique diffère un peu de celle de Meuwissen (*et al.,* 2001), du fait que la probabilité π soit estimée et non pas simplement fixée à 95% (Gustavo De los Campos, communication personnelle).

Les paramètres σ_{e}^2, μ, S_β et π sont obtenus grâce à l'échantillonneur de Gibbs (Rodriguez et De los Campos, 2013) qui utilise la méthode de Monte-Carlo par chaîne de Markov (MCMC).

1.4. Problématique

Le projet « Méthode de SAM : Evaluation de différentes méthodes de SG pour le palmier à huile » est un projet de recherche né d'une collaboration entre le CIRAD et PalmElit. L'unité AGAP du CIRAD, porteuse du projet, est spécialisée dans l'amélioration génétique et l'adaptation des plantes méditerranéennes et tropicales. PalmElit est une entreprise spécialisée dans l'amélioration génétique du palmier à huile.

En 2013, Vincent Souchard a contribué à l'établissement d'un modèle génomique sur la régularité de la production du palmier à huile. Toujours sur les mêmes données David Cros, responsable scientifique du projet, a publié en 2014 un article sur la méthodologie et la précision de la SG dans le but de prédire la BV d'individus qui n'auraient pas été testés en croisement. Le présent mémoire s'insère dans cette continuité, avec l'objectif d'affiner encore la précision de la SG et plus spécifiquement de la méthode G-BLUP. Pour cela, la piste envisagée consiste à passer d'un modèle G-BLUP univarié à un modèle G-BLUP multivarié. Les variables d'intérêt sont le nombre de régimes (NR) et leur poids moyen (PM). Notre étude vise plus précisément à répondre aux trois questions suivantes :

(i) Nous cherchons le meilleur modèle pour prédire l'AGC de parents testés en croisement. L'objectif est de valoriser au mieux les données issues de tests sur descendance.

(ii) Nous cherchons la meilleure méthode pour prédire l'AGC de parents non-testés en croisement. Cette connaissance est utile pour étendre la sélection à un ensemble de palmiers trop nombreux pour être tous testés en croisement d'une part, et d'autre part lorsque l'on réalise plusieurs cycles consécutifs de sélection (sans tester la descendance à chaque génération).

(iii) Nous nous interrogeons sur le comportement des modèles lorsque l'on réduit le nombre de marqueurs moléculaires. La finalité est de déterminer un nombre minimal et suffisant de marqueurs moléculaires afin d'ajuster au mieux le coût du génotypage.

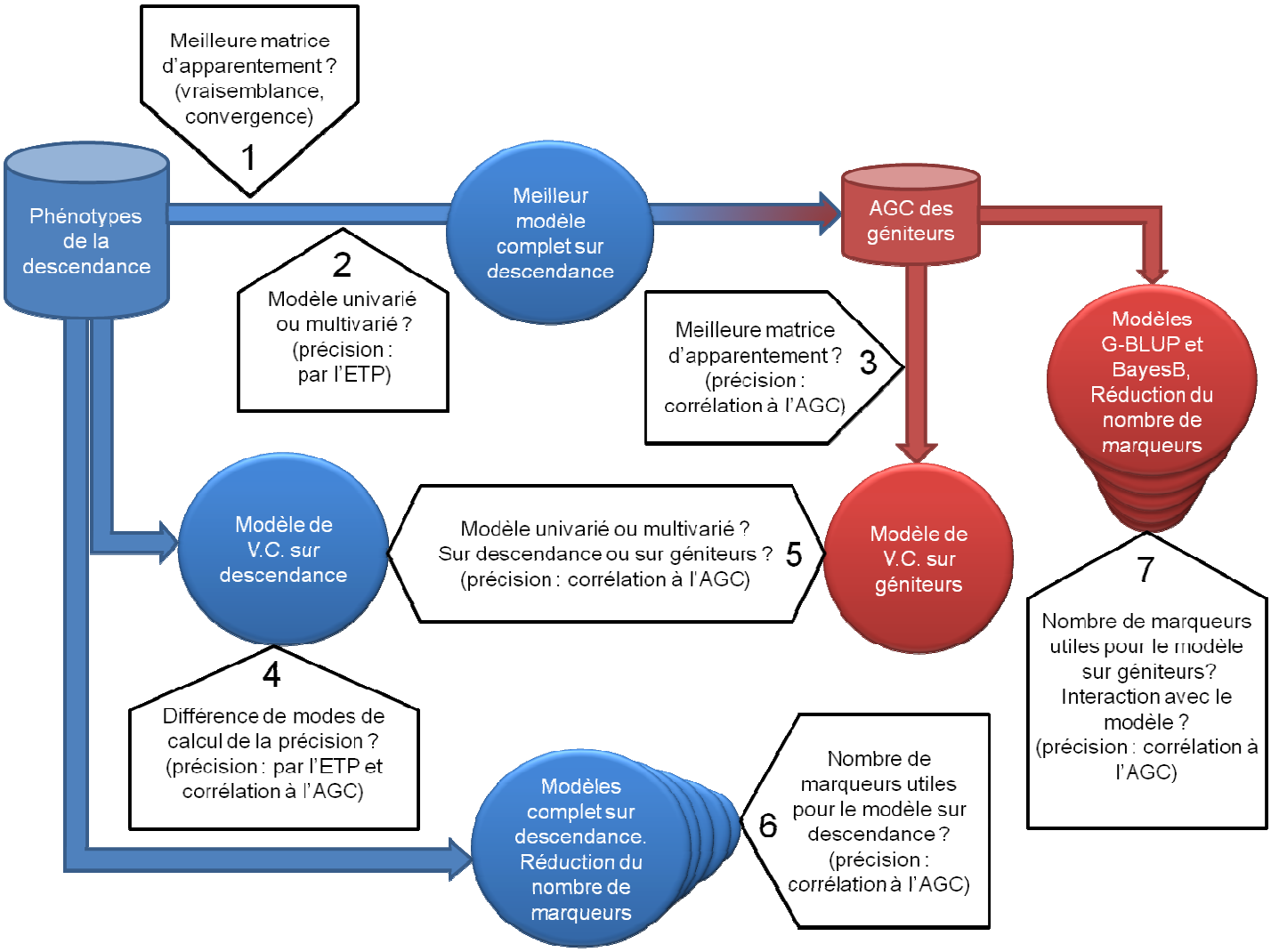


Figure 5 : Modèle d'analyse.

1.5. Modèle d'analyse

L'étude est menée en suivant le modèle d'analyse présenté en Figure 5. On utilisera deux sortes de modèles : le « modèle sur descendance » valorise les données phénotypiques des descendants hybrides observés dans les essais, et le « modèle sur géniteurs » qui est un modèle simplifié qui prédit l'AGC de géniteurs potentiels à partir de l'AGC des autres géniteurs. Pour le modèle sur descendance, l'épithète « complet » signifie que le modèle utilise l'intégralité des données phénotypiques, par opposition aux modèles utilisés en validation croisée (VC), lors de laquelle on utilise seulement les données phénotypiques d'une partie des descendants, pour prédire l'AGC des parents des autres descendants.

Le modèle sur descendance peut donc servir à la fois pour estimer la précision de sélection d'individus testés en croisement (modèle sur descendance complet) et en VC. Le modèle sur géniteurs est utilisé uniquement en VC.

Pour répondre à la première question de la problématique, nous utilisons un modèle sur descendance complet pour comparer les matrices d'apparentement moléculaire $\mathbf{G_{EM}}$, $\mathbf{G_{OF}}$, $\mathbf{G_N}$ entre elles et à la matrice d'apparentement généalogique standard **A** (Figure 5, Point 1. Nous utilisons ensuite ce même modèle pour étudier l'intérêt du passage au modèle multivarié (Figure 5, Point 2). A cette étape, la précision calculée est une précision de sélection. Le meilleur modèle complet sur descendance sert à donner les meilleures prédictions possibles des AGC des géniteurs, qui seront utilisées comme valeurs de référence pour la VC.

Pour répondre à la seconde question de la problématique, nous procédons par VC. L'utilisation de la VC simule la prédiction d'AGC de géniteurs non-testés en croisement. A cette étape, nous utilisons un autre mode de calcul de la précision : au lieu d'une précision de sélection, nous calculons une précision de prédiction qui est la corrélation entre les AGC prédites par un modèle donné et les valeurs de référence. Ici, nous pouvons utiliser les 2 types de modèles, sur descendance et sur géniteurs. Nous commençons par étudier quelle matrice d'apparentement ($\mathbf{G_{EM}}$, $\mathbf{G_{OF}}$, $\mathbf{G_N}$ voire **A)** convient le mieux au modèle sur géniteurs (Figure 5, Point 3). Pour le modèle sur descendance en VC, nous observons les différences entre les deux modes de calcul de la précision (précision de sélection *vs.* précision de prédiction) (Figure 5, Point 4). Ensuite, nous nous demandons lequel des modèles « sur descendance » ou « sur géniteurs » convient le mieux pour prédire l'AGC des parents non-testés en croisement. Nous étudions également l'intérêt du passage au modèle multivarié dans les deux cas de figure. (Figure 5, Point 5).

Finalement, nous étudions le comportement des modèles lorsque nous faisons varier le nombre de marqueurs moléculaires. Dans un premier temps, nous observons l'évolution du meilleur modèle complet sur descendance avec une densité de marqueurs décroissante (Figure 5, Point 6). Dans un second temps, nous observons l'évolution de différents modèles sur géniteurs en validation croisée : G-BLUP univarié et multivarié, mais aussi un modèle Bayésien (BayesB) univarié (Figure 5, Point 7), lorsque décroît le nombre de marqueurs. Dans tous les cas, le modèle généalogique (basé sur le pédigrée uniquement) est utilisé comme référence.

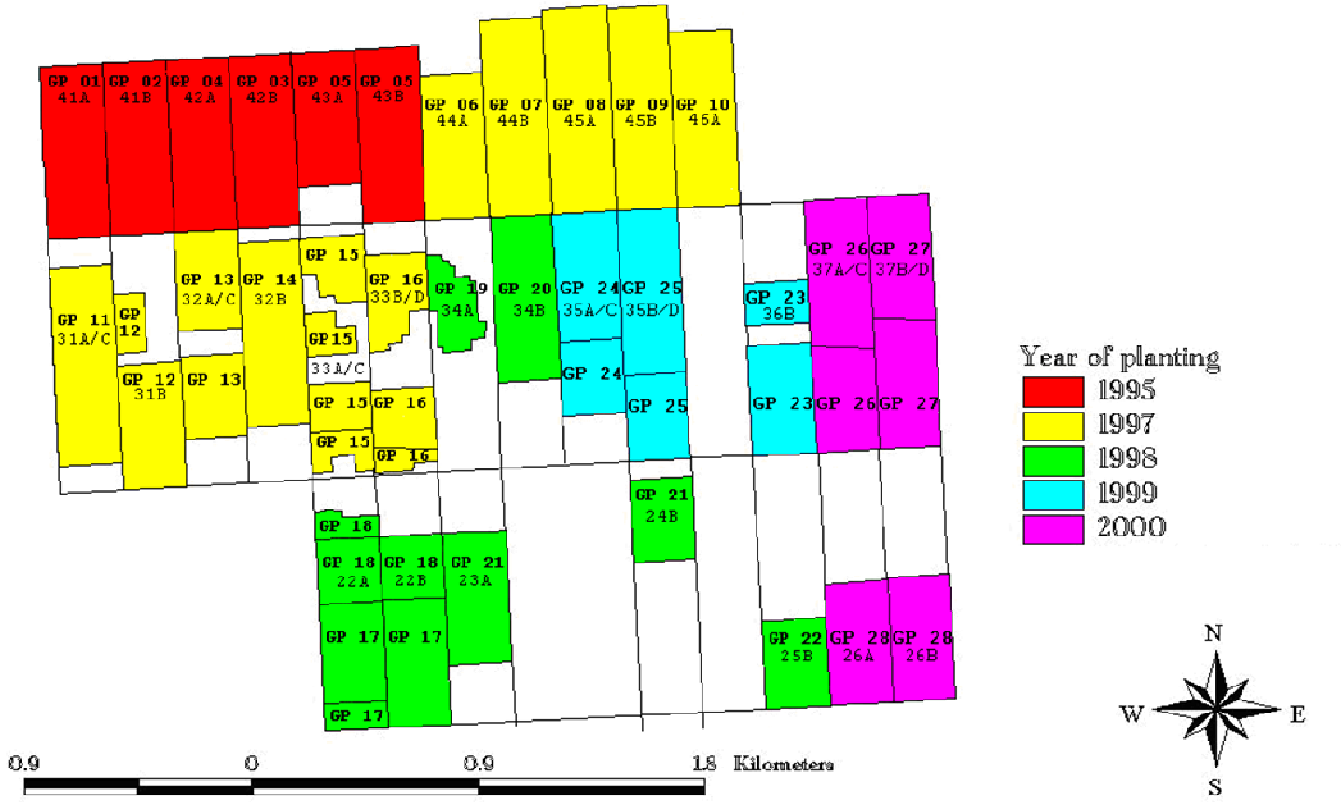


Figure 6 : Plan des 28 essais génétiques d'Aek Loba (North Sumatra, Indonésie) (les essais GP5, 19 et 22 ne sont pas utilisés). Le 26ème essai provient de Bangun Bandar (non représenté ici).

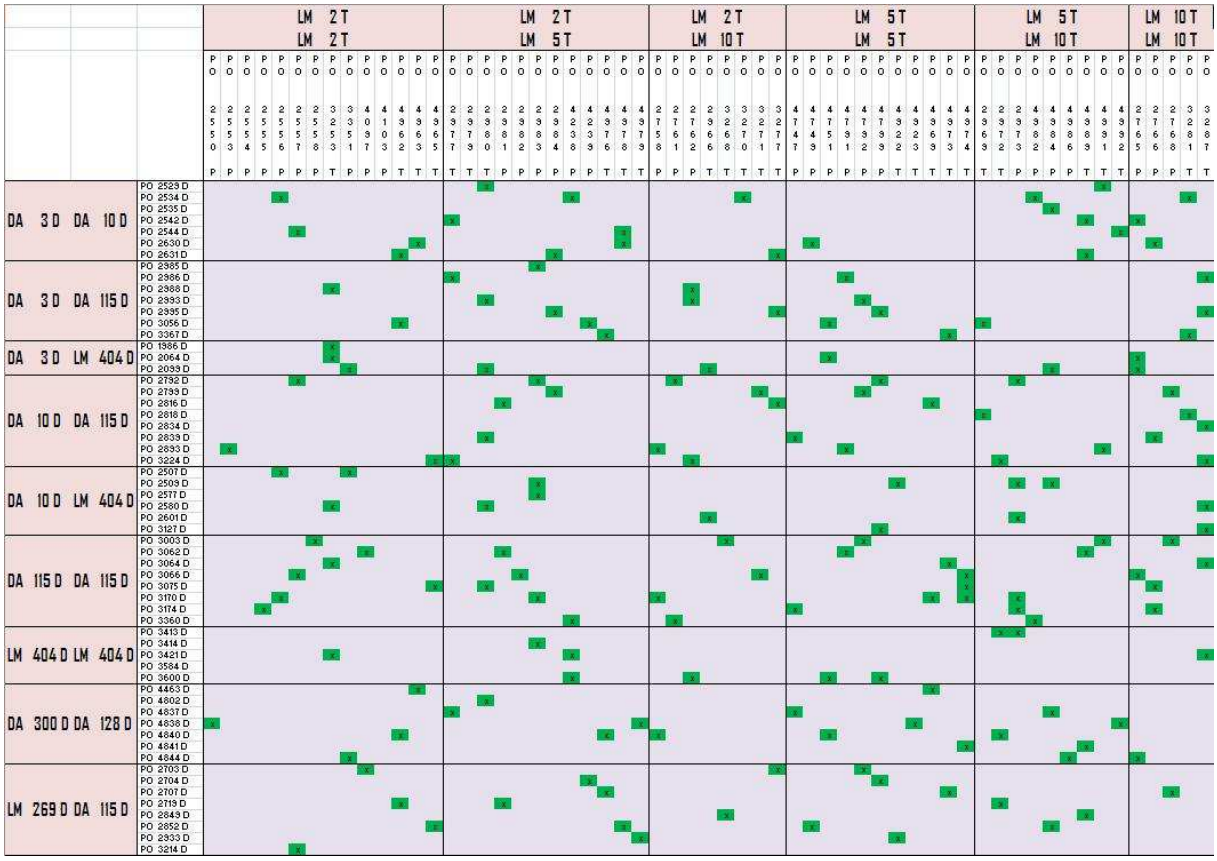


Figure 7 : Plan de croisement entre les géniteurs Deli du groupe A (à gauche) et les géniteurs La Mé du groupe B (en haut).

2. Matériel et méthodes

2.1. Recueil et manipulation des données

2.1.1. Données phénotypiques

La palmeraie expérimentale est située à Aek Loba (Sumatra du Nord, Indonésie), sa superficie avoisine 350ha et elle a été plantée entre 1995 et 2000 (Figure 6). Au total, 50140 palmiers ont été plantés. Parmi eux, 30852 sont exploitables pour l'expérimentation, c'est-à-dire vivants au moment des mesures et de type *tenera*.

La palmeraie est segmentée en 28 essais, reconnaissables sur la Figure 6 par leur identifiant « GP ». Seuls 26 essais sont valorisés dans l'étude (un essai provient de Bangun Bandar, à côté d'Aek Loba). Les essais sont organisés en blocs complets ou en lattices. Sur un essai sont testés 20 à 25 croisements. Les croisements sont réalisés de sorte à optimiser l'analyse de variance prévue. La Figure 7 montre le plan de croisements entre les parents Deli et un les parents La Mé. Chaque croisement est répété sur 5 à 6 parcelles élémentaires. Une parcelle élémentaire regroupe environ 12 palmiers tous issus du même croisement.

Les régimes sont récoltés tous les 10 jours pour les palmiers de 3 à 11 ans, ce qui permet de les compter (NR) et de les peser (PM). D'autres observations sont réalisées (hauteur, infestation à *Ganoderma*, taux d'extraction de l'huile, etc.). Les données phénotypiques ainsi collectées (NR et PM) sont les mêmes que celles utilisées par Souchard (2013) et Cros *et al.* (2014). Au total, 180872 observations de PM et NR sont disponibles.

2.1.2. Données génotypiques

Le génotypage est réalisé par des marqueurs microsatellites ou *simple sequence repeats* (SSR). La collection de données de microsatellites utilisée par Souchard (2013) et Cros *et al.* (2014) est à nouveau mobilisée. Ces microsatellites proviennent de deux banques développées par Billotte *et al.* (2005) et Tranbarger *et al.* (2012). Depuis Souchard (2013) et Cros *et al.* (2014), la collection de microsatellites a été augmentée en utilisant de nouveaux marqueurs provenant des deux banques précédemment citées, ainsi que d'une troisième banque développée par Zaki *et al.* (2012). Le gène Sh (voir 1.2.2. Déterminisme génétique des composantes du rendement) est intégré au jeu de donnée, il joue le même rôle qu'un marqueur à 2 allèles. En effet on compte dans le groupe B des hétérozygotes *tenera* et des homozygotes *pisifera*. Au final, 265 marqueurs sont disponibles pour le groupe A et 289 marqueurs sont disponibles pour le groupe B.

Les données génotypiques manquantes (environ 1.7% pour le groupe A et 2.9% pour le groupe B (Cros *et al.*, 2014)) ont été imputées en utilisant BEAGLE 3.3.2 (Browning et Browning, 2007).

2.1.3. Manipulation des données

L'ensemble de l'étude est réalisée sous R 3.0.2 (R Core Team, 2014).

2.2. Modèles sur descendance

Les modèles sur descendance permettent de prédire les AGC des géniteurs du test génétique, à partir des phénotypes des descendants.

2.2.1. Modèle mixte univarié généalogique sur descendance

Le modèle mixte univarié complet basé sur le pédigrée valorise les données d'apparentement généalogique pour prédire les AGC des géniteurs. Ecrivons-le ainsi :

$$P_{abcdeij} = \mu + essai_a + essai_a{*}rep_b + essai_a{*}exp_c + palmier_d + âge_e + AGC_{Ai} + AGC_{Bj} + ASC_{ij} + âge_e{*}ASC_{ij} + e_{abcdeij}$$

avec $P_{abcdeij}$ l'observation phénotypique (PM ou NR) n°abcdeij, réalisée sur le palmier d de parents i et j à l'âge e. En tout, 180872 observations sur les 30852 palmiers sont prises en compte dans le modèle complet (*i.e.* sans retranchement d'un jeu de VC).

a) Effets fixes

Les effets fixes de ce modèle sont :

- l'effet essai. En tout 26 essais de la palmeraie sont exploités ;
- l'effet de la répétition imbriquée dans l'essai : essai*répétition (152 niveaux). Dans le cas d'un essai organisé en blocs, ce terme correspond à l'effet bloc pour cet essai ;
- l'âge en facteur à 6 niveaux : 6 ans, 7 ans, … 11 ans.

b) Effets aléatoires non-génétiques

Les effets aléatoires non-génétiques de ce modèle sont :

- L'effet de la parcelle élémentaire imbriquée dans l'essai : essai*exp ~ N(0, $\sigma^2_{essai:exp} \times \mathbf{I}$). Ce terme correspond à l'identifiant d'une parcelle élémentaire de 12 palmiers issus d'un même croisement (3464 niveaux) ;
- L'effet palmier ~ N(0, $\sigma^2_{palmier} \times \mathbf{I}$) (30852 niveaux) ;
- Le résidu e ~ N(0, $\sigma^2_e \times \mathbf{I}$).
- L'interaction entre l'âge et la famille : âge*ASC ~ N(0, $\sigma^2_{âge*ASC} \times \mathbf{I} \otimes \mathbf{D}$) (Marie Denis, communication personnelle) (2855 niveaux) ;

c) Effets aléatoires génétiques

L'objectif du modèle est de prédire les effets génétiques additifs des géniteurs, *i.e.* leurs AGC. Les effets génétiques du modèle sont structurés (Stuber et Cockerham, 1966; Bernardo, 1996) :

- Les AGC des géniteurs du groupe A : AGC_A ~ N(0, $\sigma^2_{AGC_A} \times ½\mathbf{A_A}$) avec $\mathbf{A_A}$ la matrice d'apparentement calculée à partir du pédigrée des géniteurs du groupe A ;
- Les AGC des géniteurs du groupe B : AGC_B ~ N(0, $\sigma^2_{AGC_B} \times ½\mathbf{A_B}$) avec $\mathbf{A_B}$ la matrice d'apparentement calculée à partir du pédigrée des géniteurs du groupe B ;
- Les ASC de chaque croisement, ou effets famille : ASC ~ N(0, $\sigma^2_{ASC} \times \mathbf{D}$) avec **D** la matrice de dominance des combinaisons de géniteurs.

La préparation des matrices A et D avec les packages synbreed (Wimmer *et al.*, 2012) et nadiv (Wolak, 2012) est détaillée en Annexe A (Points (a) et (b)). La saisie du modèle sur descendance complet univarié basé sur le pédigrée avec le package ASReml (Gilmour *et al.*, 2009) est également présentée en Annexe A (Point (c)).

2.2.2. Modèles G-BLUP sur descendance

Au sein de chacun des groupes A et B, et pour l'ensemble des géniteurs testés en croisement :

- 6 géniteurs du groupe A et 25 du groupe B ne sont pas génotypés ;
- 140 géniteurs du groupe A et 131 du groupe B sont génotypés ;

Dans l'analyse classique (BLUP) des essais génétiques, tous les géniteurs sont inclus dans le modèle. Pour pouvoir faire une analyse comparable avec le G-BLUP, il faut à nouveau utiliser l'ensemble les géniteurs. Pour cela nous produisons une matrice d'apparentement combinant les données moléculaires de ceux qui ont été génotypés et les données généalogiques des autres. Cette opération se fait en utilisant la méthode de Forni *et al.* (2011) :

$$\mathbf{H^{-1}} = \mathbf{A^{-1}} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G^{-1}} - \mathbf{A_{22}^{-1}} \end{bmatrix}$$

Avec **A** la matrice d'apparentement généalogique, **G** la matrice d'apparentement moléculaire, et $\mathbf{A_{22}}$ la portion de **A** qui comprend l'information généalogique des géniteurs génotypés. On applique le modèle G-BLUP sur descendance en remplaçant $\mathbf{A_A}$ et $\mathbf{A_B}$ du modèle généalogique précédent par les matrices $\mathbf{H_A}$ et $\mathbf{H_B}$. Par contre, la matrice **D** utilisée pour prédire les ASC est toujours calculée à partir du pédigrée.

La préparation de la matrice $\mathbf{H^{-1}}$ à partir de la matrice $\mathbf{G_{EM}}$ et des pédigrées est présentée en Annexe A (Point (d)). Cette préparation utilise le package synbreed (Wimmer *et al.*, 2012).

2.2.3. Convergence des modèles sur descendance

La méthode REML mise en œuvre par le logiciel ASReml est itérative, et doit s'achever par (i) la convergence de la log-vraisemblance du REML et (ii) la convergence des estimations des paramètres de variance (Gilmour *et al.*, 2009).

Il arrive que certains modèles ne convergent pas. Afin de prévenir ce problème, nous employons la fonction « *nearest positive definite* » disponible dans le package Matrix de R (Bates et Maechler, 2014) et appelée nearPD par la suite. Cette fonction calcule la matrice définie positive la plus proche de la matrice donnée en entrée. Nous appliquons la fonction nearPD aux 4 matrices d'apparentement inversées ($\mathbf{A^{-1}}$ et $\mathbf{H^{-1}}$ calculée avec $\mathbf{G_{EM}}$, $\mathbf{G_{OF}}$, $\mathbf{G_N}$), afin éventuellement d'aider les modèles à converger. L'effet de la fonction nearPD est étudié afin de juger de la pertinence de son utilisation.

La procédure pour appliquer la fonction nearPD à une matrice $\mathbf{H^{-1}}$ est présentée en Annexe A, Point (c).

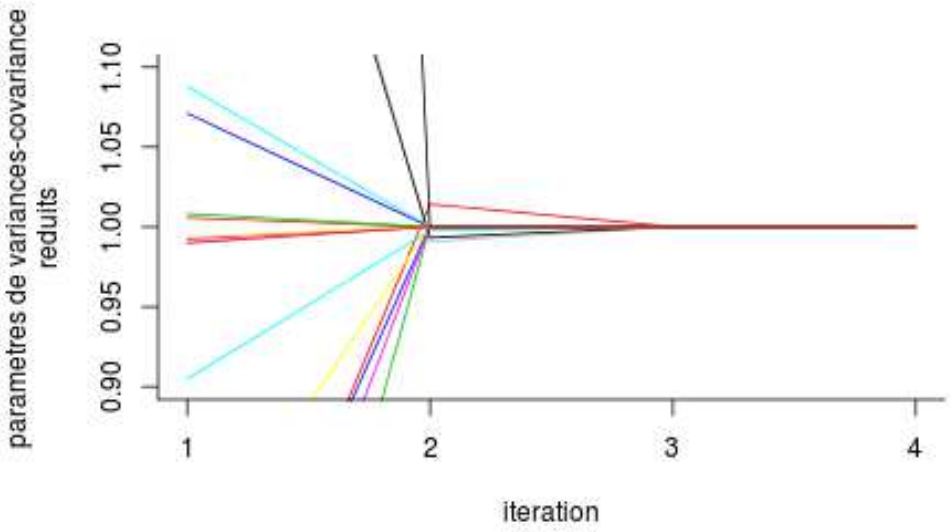


Figure 8 : Convergence en 4 itérations des paramètres initiaux de variance-covariance pour le modèle multivarié génomique complet sur descendance.

2.2.4. Modèles multivariés sur descendance

Le passage au modèle multivarié (deux caractères identifiés par les numéros 1 et 2) fait apparaître des paramètres de covariance pour les AGC :

$$\begin{bmatrix} \mathrm{AGC}_{\mathrm{A}_1} \\ \mathrm{AGC}_{\mathrm{A}_2} \end{bmatrix} \sim \mathrm{N}(0, \begin{bmatrix} \sigma^2{}_{\mathrm{AGC}_{\mathrm{A}_1}} & \mathrm{C}_{\mathrm{AGC}_{\mathrm{A}_{12}}} \\ \mathrm{C}_{\mathrm{AGC}_{\mathrm{A}_{12}}} & \sigma^2{}_{\mathrm{AGC}_{\mathrm{A}_2}} \end{bmatrix} \otimes \mathbf{A_A}) \quad \text{et} \quad \begin{bmatrix} \mathrm{AGC}_{\mathrm{B}_1} \\ \mathrm{AGC}_{\mathrm{B}_2} \end{bmatrix} \sim \mathrm{N}(0, \begin{bmatrix} \sigma^2{}_{\mathrm{AGC}_{\mathrm{B}_1}} & \mathrm{C}_{\mathrm{AGC}_{\mathrm{B}_{12}}} \\ \mathrm{C}_{\mathrm{AGC}_{\mathrm{B}_{12}}} & \sigma^2{}_{\mathrm{AGC}_{\mathrm{B}_2}} \end{bmatrix} \otimes \mathbf{A_B})$$

Pour le modèle génomique, on remplacera **A** par **H**. Des paramètres de covariances apparaissent également pour les résidus :

$$\begin{bmatrix} \mathrm{e}_1 \\ \mathrm{e}_2 \end{bmatrix} \sim \mathrm{N}(0, \begin{bmatrix} \sigma^2{}_{\mathrm{e}_1} & \mathrm{C}_{\mathrm{e}_{12}} \\ \mathrm{C}_{\mathrm{e}_{12}} & \sigma^2{}_{\mathrm{e}_2} \end{bmatrix} \otimes \mathbf{I})$$

Pour réaliser le modèle multivarié avec le logiciel ASReml il est nécessaire de proposer une valeur initiale de variance et de covariance pour tous les effets aléatoires (Gilmour *et al.*, 2009). Cela revient à proposer $\sigma^2{}_{\mathrm{AGC}_{\mathrm{A}_1}}$, $\mathrm{C}_{\mathrm{AGC}_{\mathrm{A}_{12}}}$, $\sigma^2{}_{\mathrm{AGC}_{\mathrm{A}_2}}$, $\sigma^2{}_{\mathrm{AGC}_{\mathrm{B}_1}}$, $\mathrm{C}_{\mathrm{AGC}_{\mathrm{B}_{12}}}$, $\sigma^2{}_{\mathrm{AGC}_{\mathrm{B}_2}}$, $\sigma^2{}_{\mathrm{e}_1}$, $\mathrm{C}_{\mathrm{e}_{12}}$, $\sigma^2{}_{\mathrm{e}_2}$ mais aussi des variances pour les autres effets aléatoires, $\sigma^2{}_{\mathrm{palmier}_1}$, $\sigma^2{}_{\mathrm{palmier}_2}$, $\sigma^2{}_{\mathrm{essai*exp}_1}$, $\sigma^2{}_{\mathrm{essai*exp}_2}$, $\sigma^2{}_{\mathrm{ASC}_1}$, $\sigma^2{}_{\mathrm{ASC}_2}$, $\sigma^2{}_{\mathrm{âge*ASC}_1}$ et $\sigma^2{}_{\mathrm{âge*ASC}_2}$, soit 17 paramètres au total. Dans un premier temps, les paramètres de variance ont été estimés par REML au moyen de modèles univariés. Les paramètres de covariances ont été calculés comme suit :

$$\mathrm{C}_{\theta_{12}} = \mathrm{C}_{12} \times \sqrt{\sigma^2_{\theta_1} \times \sigma^2_{\theta_2}}$$

Avec $\mathrm{C}_{12} = \mathrm{C}_{\mathrm{PM,NR}} = -0.686$ la corrélation phénotypique entre PM et NR dans l'ensemble du jeu de données. Les paramètres de variance-covariance estimés par le modèle multivarié ont ensuite été récupérés, et utilisés eux-mêmes comme des paramètres initiaux. Ce processus a été opéré en boucle jusqu'à ce que les paramètres, arrondis à 10^{-3}, ne diffèrent plus d'une itération à l'autre.

Cette méthode a été utilisée pour le modèle généalogique complet sur descendance et pour le modèle génomique complet sur descendance. L'évolution des valeurs réduites (divisées par leur valeur finale) de ces paramètres à chaque itération est présentée pour le modèle génomique en Figure 8. On se rend compte qu'elles convergent en 4 itérations vers ce qui semble être un état stable. La convergence de ces paramètres pour le modèle généalogique se déroule également en 4 itérations. Pour les valeurs de variance-covariance de la première itération, les calculs pour les modèles basés sur le pédigrée et sur les marqueurs sont calculés en 12.9h et 14.6h respectivement. Avec les valeurs de variance-covariance obtenues à la dernière itération, ces modèles sont calculés en 15.6min et 18.2min respectivement.

Les paramètres de variance-covariance finaux sont utilisés pour initialiser tous les modèles sur descendance, respectivement généalogiques et génomiques, qu'ils soient complets ou utilisés en VC. La saisie du modèle sur descendance complet multivarié valorisant l'information génomique, avec le package ASReml (Gilmour *et al.*, 2009), est détaillée en Annexe A (Point (e)).

2.2.5. Critères de vraisemblance

Il n'est pas possible de comparer la précision de sélection calculée par la formule de la VEP (Cf 1.2.6. Erreur de prédiction) de modèles dont la matrice d'apparentement diffère, car les populations de base auxquelles se réfèrent ces précisions ne sont pas les mêmes (Andrés Legarra, communication personnelle). Or, changer de matrice d'apparentement d'un modèle n'impacte ni son nombre de paramètres à estimer, ni le nombre d'observations. Pour cette raison, comparer la vraisemblance (ou la déviance, *i.e.* la -2 log(vraisemblance)) de deux modèles dont seule change la matrice d'apparentement est strictement équivalent à comparer leurs critères de vraisemblance pénalisés (critère d'information d'Akaïké (AIC) et critère d'information Bayésien (BIC)). De ce fait, nous pouvons comparer directement la déviance des modèles lorsque la matrice d'apparentement diffère. La comparaison des vraisemblances est donc employée lorsque l'on cherche la meilleure matrice d'apparentement pour le modèle complet sur descendance (Point 1 du modèle d'analyse, Cf Figure 5, Point 1).

2.2.6. Précision de sélection

La précision de sélection des modèles complets sur descendance est calculée en utilisant la variance d'erreur de prédiction (Cf 1.2.6. Erreur de prédiction). Nous soulignons qu'il est identique de calculer la précision de sélection à partir de la variance d'erreur de prédiction de la BV, avec la variance additive au dénominateur, ou à partir de la variance d'erreur de prédiction de l'AGC, avec la variance de l'AGC au dénominateur :

$$r_x = \sqrt{1 - \frac{VEP_{AGC}}{\Gamma_{xx}}} = \sqrt{1 - \frac{VEP_{AGC}}{½(1 + F_x)\sigma_g^2}}$$

Nous calculons la précision de sélection des modèles sur descendances utilisant la meilleure matrice d'apparentement. La consanguinité F_i est lue dans cette meilleure matrice d'apparentement. La variance d'AGC σ^2_g est estimée grâce aux modèles. La VEP est propre aux modèles.

Dans un premier temps (Figure 5, point 2) la précision de sélection est calculée avec les modèles complets univariés puis avec le modèle multivarié, pour les variables PM et NR, et pour tous les géniteurs testés en croisement et génotypés (140 pour le groupe A et 131 pour le groupe B). Les précisions de sélection des modèles univariés et du modèle multivarié sont ensuite comparées avec un test t de Student sur données appariées, pour chaque combinaison de groupe (A ou B) et de variables (PM ou NR). Les 4 p-values ainsi obtenues sont ajustées avec une correction de Bonferroni. Ces outils statistiques sont disponibles dans le package stats (R Core Team, 2014).

Dans un second temps (Cf Figure 5, Point 4), nous comparons la précision de sélection et la précision de prédiction du modèle sur descendance en VC. Ces précisions sont calculées pour les géniteurs composant le jeu de validation (parmi les 131 géniteurs du groupe A et les 131 géniteurs du groupe B utilisés en VC).

2.3. Modèles sur géniteurs

Après avoir estimé l'AGC des géniteurs avec le meilleur modèle complet sur descendance, nous utilisons l'AGC d'un groupe de géniteurs (jeu d'apprentissage) pour prédire l'AGC d'un autre groupe de géniteurs (jeu de validation). Les modèles sur géniteurs servent à prédire l'AGC de géniteurs non-testés en croisement, et pour cette raison nous les utiliserons toujours en validation croisée (VC) (Cf 2.4. Validation croisée).

Les jeux de VC utilisés proviennent de l'étude de Cros *et al.* (2014). Ils comprennent les géniteurs génotypés et testés en croisement. Cependant les 9 géniteurs Angola du groupe A ont été écartés car ils sont trop peu nombreux et fortement consanguins (issus de l'autofécondation d'un seul palmier). Les jeux de validations comprennent donc les 131 géniteurs Deli du groupe A et les 131 géniteurs du groupe B (La Mé et Yangambi).

2.3.1. Modèle mixte univarié généalogique sur géniteurs

Le modèle mixte sur géniteurs prédit les AGC des géniteurs ($\widehat{\mathrm{AGC}}$) à partir des estimations d'AGC obtenues par le meilleur modèle complet sur descendance. Ces modèles s'appliquent aux groupes A et B de géniteurs indépendamment l'un de l'autre :

$$\mathrm{AGC}_{\mathrm{A_i}} = \mu_{\mathrm{A}} + \widehat{\mathrm{AGC}}_{\mathrm{A_i}} + \mathrm{e}_{\mathrm{A_i}} \qquad \text{et} \qquad \mathrm{AGC}_{\mathrm{B_i}} = \mu_{\mathrm{B}} + \widehat{\mathrm{AGC}}_{\mathrm{B_i}} + \mathrm{e}_{\mathrm{B_i}}$$

Les AGC prédites par le modèle sur géniteur sont considérées comme des effets aléatoires, telles que pour le modèle généalogique :

$$\widehat{\mathrm{AGC}}_{\mathrm{A}} \sim \mathrm{N}\left(0, \sigma^2_{\widehat{\mathrm{AGC}}_{\mathrm{A}}} \times \tfrac{1}{2}\mathbf{A_{22_A}}\right) \qquad \text{et} \qquad \widehat{\mathrm{AGC}}_{\mathrm{B}} \sim \mathrm{N}(0, \sigma^2_{\widehat{\mathrm{AGC}}_{\mathrm{B}}} \times \tfrac{1}{2}\mathbf{A_{22_B}})$$

Où la matrice $\mathbf{A_{22_A}}$ contient l'information généalogique des 131 géniteurs Deli du groupe A et la matrice $\mathbf{A_{22_B}}$ contient l'information généalogique des 131 géniteurs du groupe B.

2.3.2. Modèles G-BLUP sur géniteurs

On passe au modèle génomique en remplaçant $\mathbf{A_{22_A}}$ par $\mathbf{G_A}$ et $\mathbf{A_{22_B}}$ par $\mathbf{G_B}$ où la matrices $\mathbf{G_A}$ contient l'information génomique des 131 géniteurs Deli du groupe A et la matrices $\mathbf{G_B}$ contient l'information génomique des 131 géniteurs du groupe B. A nouveau, nous nous demandons laquelle des matrices **G** ($\mathbf{G_{EM}}$, $\mathbf{G_{OF}}$ ou $\mathbf{G_N}$) est la plus adaptée.

2.3.3. Convergence des modèles sur géniteurs

Comme le modèle sur descendance (Cf 2.2.4. Convergence des modèles), il arrive que les modèles sur géniteurs ne convergent pas. Nous nous interrogeons à nouveau sur l'utilité de la fonction nearPD appliquée aux matrices inversées fournies à la fonction ASReml.

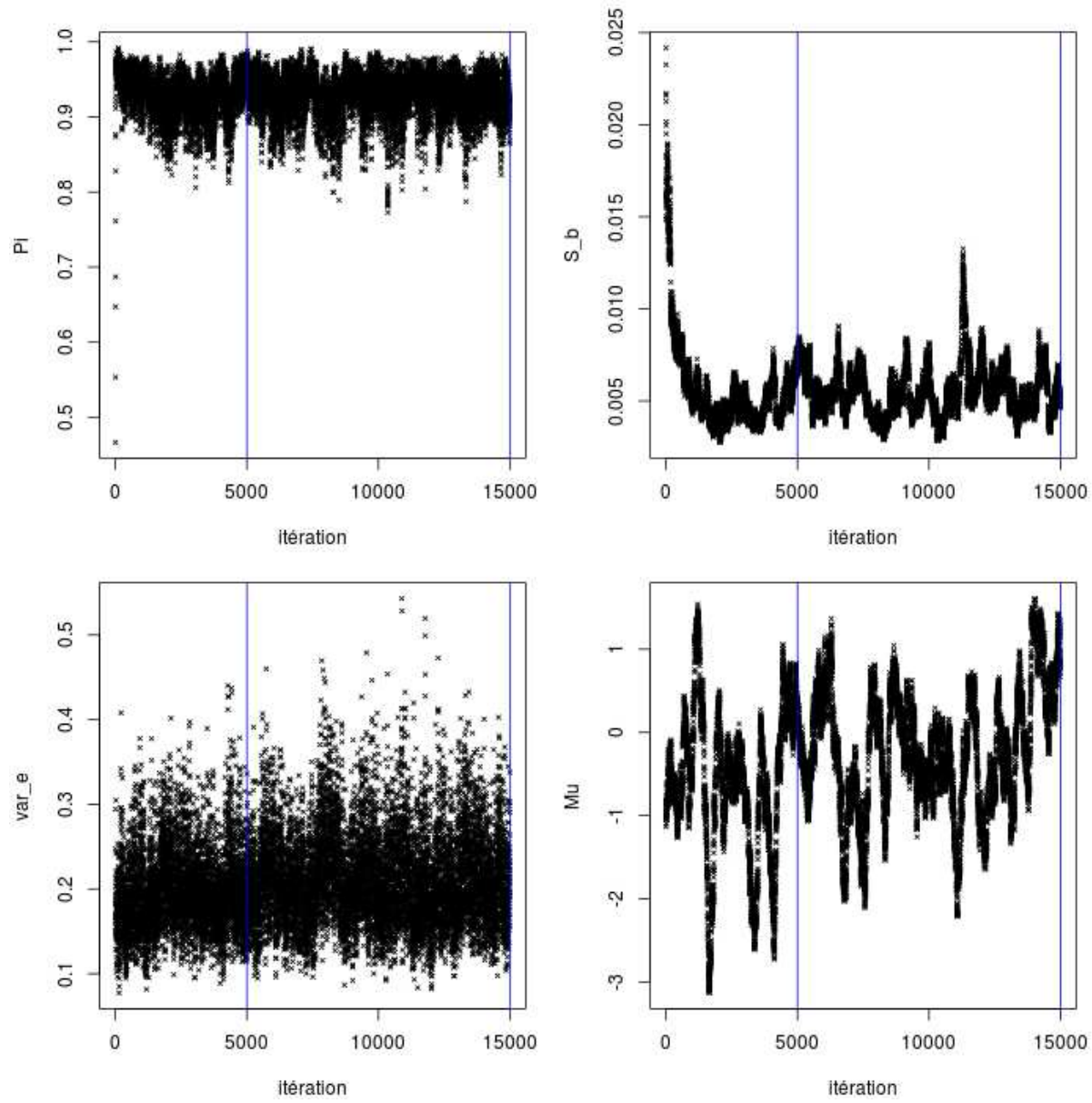


Figure 9 : Valeurs prises par les paramètres (π, S_b, σ_e^2 et μ) à chaque itération, lors de l'étape de paramétrage du modèle Bayésien (ici, pour l'une des répétitions de la VC de type « Family », appliquée aux géniteurs du groupe B, pour la variable ABW et avec 289 marqueurs moléculaires).

Seules les itérations de 5000 (burn-in) à 15000 (limites bleues verticales) seront retenues pour l'estimation des paramètres.

2.3.4. Modèles multivariés sur géniteurs

Le passage au modèle multivarié fait apparaître 3 termes de variance-covariance génétique ($\sigma^2_{\widehat{AGC}_{A_1}}$, $\sigma^2_{\widehat{AGC}_{A_2}}$, $C_{\widehat{AGC}_{A_{12}}}$ pour le groupe A, idem pour le groupe B) et 3 termes de variance-covariance résiduelle ($\sigma^2_{e_{A_1}}$, $\sigma^2_{e_{A_2}}$, $C_{e_{A_{12}}}$ pour le groupe A, idem pour le groupe B). L'initialisation de ces paramètres a été entreprise de la même façon que pour le modèle complet sur descendance (Cf 2.2.3. Modèles multivariés), cependant nous constatons que cet algorithme ne conduit pas à un équilibre stable. En effet, si l'on recommence plusieurs itérations de modèles recyclant les paramètres de variance-covariance, ils peuvent se stabiliser à 10^{-3} près pendant quelques itérations puis varier brusquement l'itération d'après. Nous prenons donc la première valeur de paramètres de variance-covariance pour laquelle les paramètres ne varient pas de 10^{-3} d'une itération sur l'autre.

Un exemple de saisie de modèle sur géniteurs multivarié génomique avec ASReml (Gilmour *et al.*, 2009) est présenté en Annexe A, Point (f).

2.3.5. Modèle BayesB sur géniteurs

Le modèle BayesB sur géniteurs se présente ainsi :

$$\hat{g}_{A_i} = \mu_A + \mathbf{Z_{A_i}}\mathbf{b_A} + e_{A_i} \qquad \text{et} \qquad \hat{g}_{B_i} = \mu_B + \mathbf{Z_{B_i}}\mathbf{b_B} + e_{B_i}$$

Le vecteur $\mathbf{Z_{A_i}}$ contient le génotype de l'individu i du groupe A (idem pour B). Les vecteurs de paramètres $\mathbf{b_A}$ et $\mathbf{b_B}$ contiennent les effets aux allèles des groupes A et B respectivement.

Le modèle est paramétré en observant la manière dont les estimations des paramètres σ^2_e, μ, S_β et π évoluent au fil des itérations de l'échantillonnage de Gibbs. Nous cherchons à déterminer combien d'itérations sont nécessaires pour que l'échantillonneur de Gibbs ait convergé. Pour chacun de ces 4 paramètres, nous avons observé les 15000 premières itérations de l'échantillonneur de Gibbs, sans observer de période de chauffe (*burn-in*), et avec un pas (*thin*) de 1, pour les 11 jeux d'entraînement (Cf 2.4.1. Jeux de VC) de chacun des groupes (la population Deli du groupe A et le groupe B), pour chaque variable (PM et NR), et en utilisant l'ensemble des marqueurs (265 et 289 pour les groupes Deli et B respectivement) ou alors seulement 10 marqueurs tirés au hasard. En tout, 88 graphiques ont ainsi été vérifiés pour chacun des 4 paramètres.

Nous concluons de cette observation qu'à la 5000ème itération, l'échantillonneur de Gibbs a convergé. Le comportement de l'échantillonneur de Gibbs pour chacun des 4 paramètres est donné à titre illustratif en Figure 9, pour l'un des 88 séries de graphiques analysés. A l'issue de cette analyse, nous décidons donc d'observer une période de chauffe de 5000 itérations, suffisante dans l'ensemble des cas. Nous décidons d'utiliser les 10000 itérations qui succèdent à la période de chauffe, avec un pas de 10, pour estimer les paramètres σ^2_e, μ, S_β et π. Nous valorisons ainsi 1000 itérations de l'échantillonneur de Gibbs.

La saisie du modèle BayesB avec le package BGLR (Rodriguez et De los Campos, 2013), en VC, est illustrée en Annexe A, Point (g).

2.4. Validation croisée

2.4.1. Jeux de VC

Les jeux de VC ont été réalisés par Cros *et al.* (2014). Ils consistent en une segmentation, groupe par groupe, des 131 géniteurs Deli du groupe A et des 131 géniteurs du groupe B. Cette segmentation est opérée selon 3 techniques qui permettent de faire varier l'apparentement entre les géniteurs du jeu d'apprentissage et ceux du jeu de validation, facteur connu pour avoir une forte influence sur la précision de la SG (Cros *et al.*, 2014).

a) Cluster

La technique dite « Cluster » vise à minimiser l'apparentement entre les géniteurs des jeux d'apprentissage et de validation. Au sein d'un groupe donné (Deli ou B), les géniteurs sont séparés par la méthode des k-means pour donner 5 clusters. La taille de chacun des 5 clusters varie entre 10.7% et 35.9% de la taille du groupe Deli, et entre 9.2% et 39.7% de la taille du groupe B.

b) Family

La technique dite « Family » vise à assurer un bon apparentement entre les 5 segments du groupe (Deli ou B). Les géniteurs de chaque famille de pleins-frères sont répartis dans les 5 segments, ainsi chacun des 5 segments contient des géniteurs de chaque famille. La taille des 5 segments varie de 19.1% à 21.4% du groupe Deli, et de 17.6% à 23.7% du groupe B.

c) CDmeans

Rincent *et al.* (2012) proposent une méthode pour optimiser l'apparentement entre les un jeu d'apprentissage et de validation. L'objectif de ces auteurs est de fournir une base méthodologique pour prédire au mieux l'aptitude de géniteurs non-testés en croisement. Un unique jeu de validation est produit pour le groupe Deli (14.5% du groupe) et pour le groupe B (19.8% du groupe).

2.4.2. Application de la VC

a) Modèle sur descendance

On retire à l'ensemble du jeu de données les croisements impliquant les parents (Deli ou B) qui sont dans le jeu de validation. Cela revient à soustraire de 4.4% à 34.2% des 180 872 observations du test génétique, soit en moyenne 16.0% de ces observations. Les données restantes sont utilisées pour calibrer le modèle et prédire l'AGC des parents non-testés.

b) Modèle sur géniteurs

Les modèles sur géniteurs sont toujours utilisés en VC (Cf 2.3. Modèles sur géniteurs).

2.4.3. Précision de prédiction

Lorsque nous réalisons une VC, nous utilisons la précision de prédiction pour juger de la qualité du modèle. Nous soulignons que $r_{\mathrm{AGC1,AGC2}} = r_{\mathrm{BV1,BV2}}$, un facteur ½ ne modifie pas les résultats d'un calcul de corrélation. Nous définissons donc la précision de prédiction comme la corrélation entre les AGC des géniteurs du jeu de validation prédites par le modèle ($\widehat{\mathrm{AGC}}$), et les AGC des géniteurs du jeu de validation prédites par le meilleur modèle complet sur descendance.

La précision de prédiction et la précision de sélection du modèle sur descendance sont comparées en VC (Cf Figure 5, Point 4). La précision de prédiction est utilisée pour évaluer les modèles qui permettront de répondre aux Points 3, 5, 6 et 7 du modèle d'analyse
(Figure 5).

2.5. Réduction du nombre de marqueurs

2.5.1. Principe général

Des ensembles de marqueurs sont tirés au sort parmi les 265 marqueurs de la population Deli ou parmi les 289 marqueurs du groupe B. On fait ainsi varier le nombre de marqueurs depuis 10 jusqu'au maximum de marqueurs, avec un pas de 10 marqueurs (*i.e.* {10, 20, 30, …, 260, 265} pour la population Deli, et {10, 20, 30, …, 280, 289} pour le groupe B). Pour chaque niveau de nombre de marqueurs, on réalise 5 répétitions de tirage au sort.

2.5.2. Application au meilleur modèle complet sur descendance

Lorsque le nombre de marqueurs moléculaires de l'un des deux groupes (A ou B) varie, on conserve l'intégralité des marqueurs moléculaires de l'autre groupe. La précision de prédiction est calculée en utilisant l'ensemble des géniteurs pour lesquels on dispose d'un génotype et qui participent au test génétique (140 géniteurs pour le groupe A et 131 géniteurs pour le groupe B). La précision des modèles génomiques est moyennée sur les 5 répétitions de tirage au sort, et comparée pour chaque groupe (A et B) et pour chaque variable (PM et NR) à la précision de prédiction du modèle généalogique (dont la précision ne dépend donc pas de la densité de marqueurs).

2.5.3. Application aux modèles sur géniteurs en VC

On applique à chaque tirage de marqueurs les modèles sur géniteurs G-BLUP univarié, G-BLUP multivarié, et BayesB. La précision de prédiction de ces modèles, en VC, est considérée. Nous comparons l'évolution de la précision de ces 3 modèles génomiques, en comparaison avec celle du modèle BLUP univarié basé sur le pédigrée. Tout d'abord, les précisions sont moyennées pour les 5 répétitions de tirage au sort à chaque densité. Puis les résultats obtenus avec chacun des 4 modèles sont comparés par un test HSD (*Honest significant difference*) de Tukey, les 11 itérations de VC constituant les répétitions. Le test est reproduit pour chaque combinaison de groupe (Deli et B) et de variable (PM et NR), et aux densités {10, 30, 70, 100, 150, 200, 250, 265 ou 289}. Nous utilisons pour cela le package agricolae (De Mendiburu, 2012).

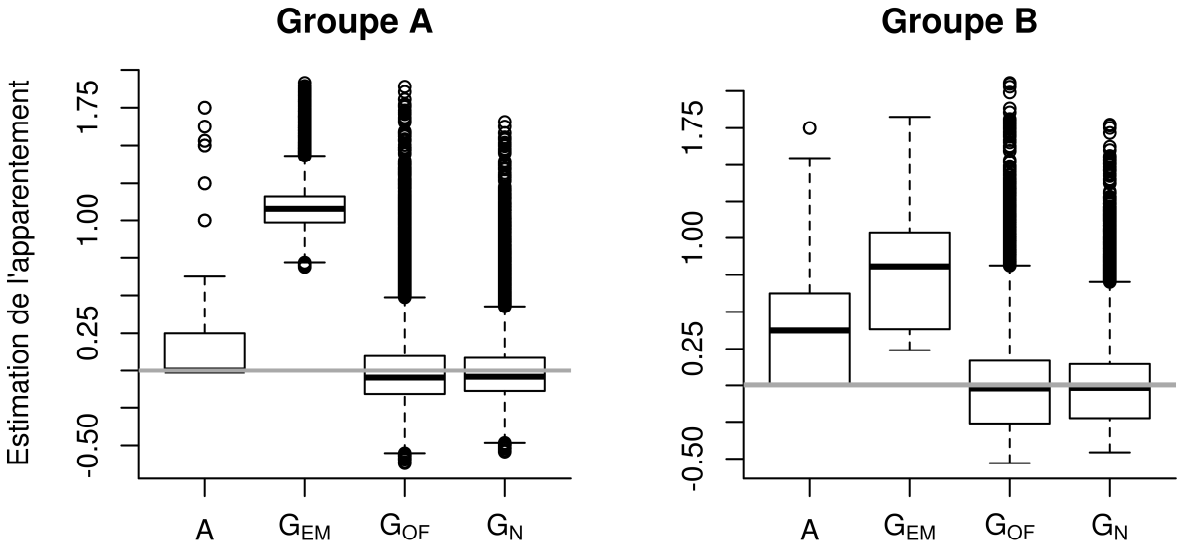


Figure 10 : Boxplot du contenu des matrices d'apparentement généalogique : ***A*** *et moléculaires :* $\boldsymbol{G_{EM}}$, $\boldsymbol{G_{OF}}$, $\boldsymbol{G_N}$ *pour les 140 géniteurs du groupe A et les 131 géniteurs du groupe B qui à la fois sont génotypés et dont on connait le phénotype.*

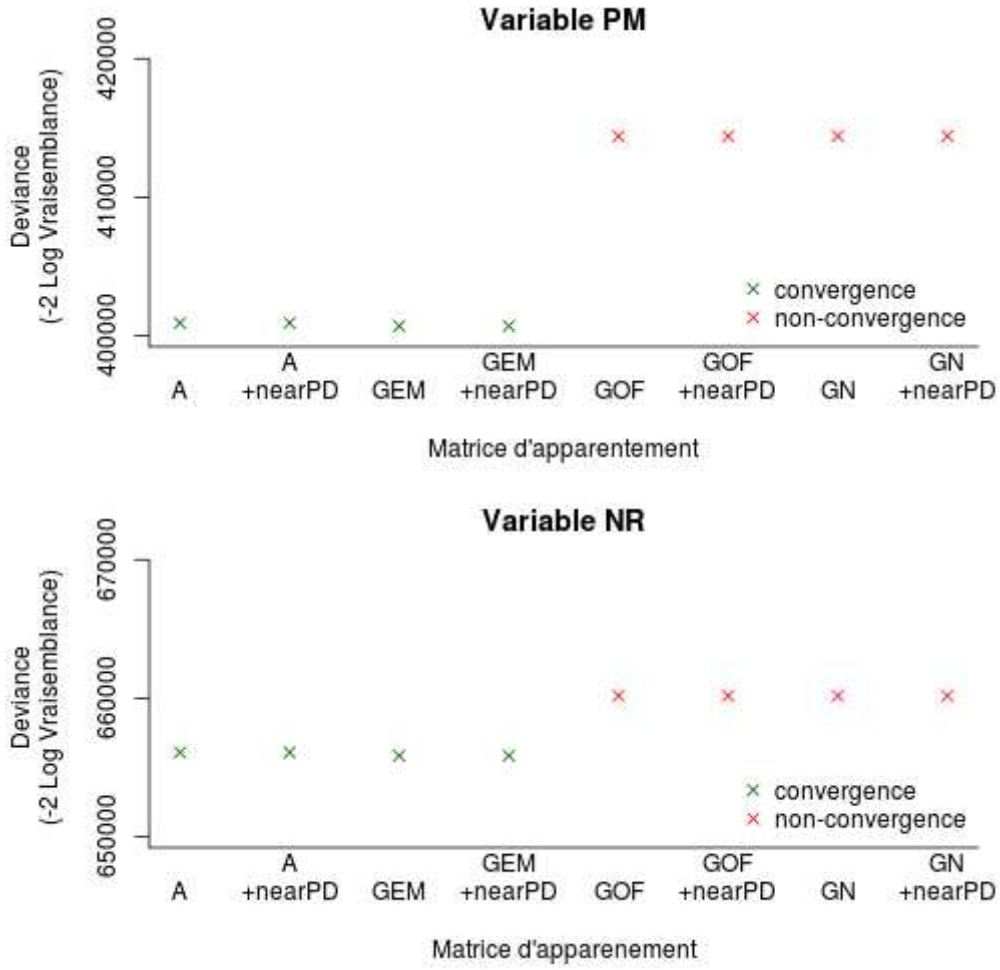


Figure 11 : Déviance (-2 LogVraisemblance) du modèle univarié complet sur descendance en fonction de la matrice d'apparentement, pour les variables PM et NR successivement. Un symbole rouge précise que le modèle n'a pas convergé. L'indication « nearPD » signifie que la fonction « nearest positive definite » a été appliquée.

3. Résultats

3.1. Optimiser le modèle de SG pour des individus testés en croisement

3.1.1. Choix de la matrice d'apparentement

Le contenu des matrices d'apparentement généalogique (**A**) et moléculaire ($\mathbf{G_{EM}}$, $\mathbf{G_{OF}}$, $\mathbf{G_N}$) est présenté en Figure 10, pour chaque groupe. Les distributions des valeurs rencontrées pour un groupe donné au sein des matrices de VanRaden $\mathbf{G_{OF}}$ et $\mathbf{G_N}$ se ressemblent beaucoup. Surprenamment, ces distributions sont presque centrées sur 0, et leurs médianes sont négatives. L'apparentement devant être compris dans l'intervalle [0 ; 2], ces matrices manquent donc de pertinence du point de vue de la génétique quantitative. Le contenu des matrices $\mathbf{G_{EM}}$ et **A** appartient bien à l'intervalle [0 ; 2]. Pour chaque groupe A et B, la distribution du contenu de $\mathbf{G_{EM}}$ est plus élevée que la distribution du contenu de **A**. En effet, le contenu de $\mathbf{G_{EM}}$ varie de 0.689 à 1.915, avec une moyenne de 1.084, pour le groupe A et de 0.241 à 1.825, avec une moyenne de 0.765, pour le groupe B; alors que le contenu de A varie de 0 à 1.75 pour les deux groupes, avec une moyenne de 0.138 pour le groupe A et 0.346 pour le groupe B. $\mathbf{G_{EM}}$ ne contient aucun apparentement nul alors que **A** en contient 73.6% pour le groupe A et 42.9% pour le groupe B. Bien que la distribution du contenu de **A** soit plus basse pour le groupe A que pour le groupe B, la distribution du contenu de $\mathbf{G_{EM}}$ est plus haute pour le groupe A que pour le groupe B, ce qui fait que le contraste entre **A** et $\mathbf{G_{EM}}$ est plus important pour le groupe A.

Pour les deux variables PM et NR, les modèles univariés sur descendance basés sur les matrices **A** et $\mathbf{G_{EM}}$ convergent tandis que les modèles univariés sur descendance basés sur les matrices $\mathbf{G_{OF}}$ et $\mathbf{G_N}$ ne convergent pas. Les modèles univariés mobilisant ces deux dernières matrices sont associés à de faibles vraisemblances (donc de fortes déviances) (Figure 11).

Les matrices $\mathbf{G_{OF}}$ et $\mathbf{G_N}$ n'étant ni pertinentes d'un point de vue de la génétique quantitative ni performantes d'un point de vue statistique sont donc écartées.

L'utilisation de la fonction nearPD n'apporte strictement aucune amélioration ni en termes de convergence ni en termes de vraisemblance au modèle univarié complet sur descendance (Figure 11), elle ne sera donc pas utilisée par la suite.

La vraisemblance des modèles complets sur descendance est meilleure (déviance plus faible) lorsque l'on utilise l'apparentement moléculaire (matrice $\mathbf{G_{EM}}$) que lorsque l'on utilise l'apparentement généalogique (Tableau 4). Cette observation est valable pour les modèles univariés prédictifs de PM et de NR, et pour le modèle multivarié les prédisant simultanément. L'utilisation de données moléculaires dans le modèle d'analyse des tests génétiques, au lieu des données généalogiques utilisées classiquement, permet donc d'obtenir des résultats de meilleure qualité.

Nous utiliserons par la suite la matrice $\mathbf{G_{EM}}$ pour les modèles G-BLUP sur descendance, car ainsi les modèles convergent et leur vraisemblance est maximisée.

Tableau 4 : Comparaison de la déviance entre les modèles génomiques (matrice $\boldsymbol{G_{EM}}$) et les modèles basés sur le pédigrée, pour les modèles complets sur descendance.

Modèle	Généalogique	Comparaison	Génomique
PM	401355,2	>	401211,2
NR	656174,2	>	656032,6
PM + NR	1053557,2	>	1053287,4

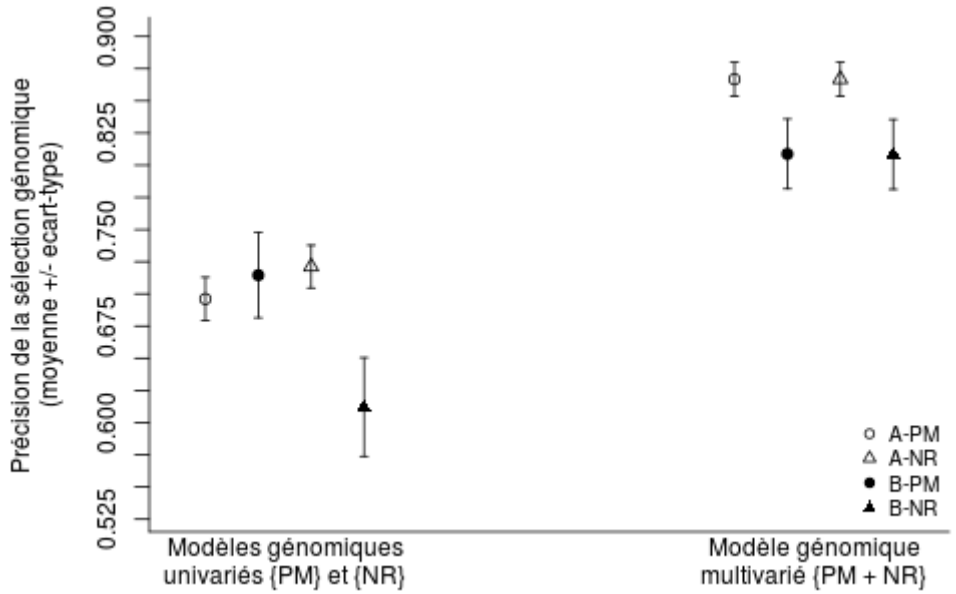


Figure 12 : Comparaison de la précision des modèles complets sur descendance G-BLUP univariés avec la précision du modèle multivarié pour PM et NR. Les précisions sont calculées sur les 140 géniteurs du groupe A et les 131 géniteurs du groupe B.

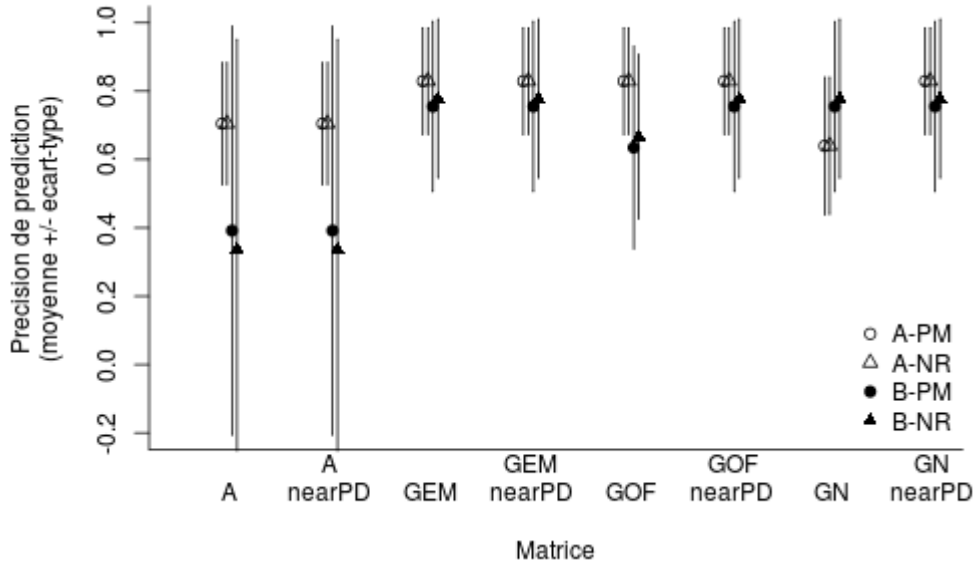


Figure 13 : Précision des modèles univariés sur géniteurs en fonction de la matrice d'apparentement utilisée. Les précisions sont calculées sur les 131 géniteurs Deli du groupe A les 131 géniteurs du groupe B. L'indication « nearPD » signifie que la fonction « nearest positive definite » a été appliquée.

3.1.2. Comparaison des modèles univariés et multivarié

La matrice d'apparentement $\mathbf{G_{EM}}$ étant fixée nous pouvons comparer les précisions de sélection des modèles complets sur descendance G-BLUP univariés et G-BLUP multivarié. Nous remarquons que pour les deux variables PM et NR et pour les deux populations A et B, le modèle multivarié donne les meilleures précisions de sélection (Figure 12). Le gain de précision de sélection lors du passage au modèle multivarié est significatif pour chaque combinaison de variable et de population ($p<10^{-100}$, test t de Student sur données appariées, n=131 géniteurs pour le groupe A et n=140 géniteurs pour le groupe B, correction de Bonferroni). Il apparait qu'entre les deux caractères, la moins bonne des précisions (PM pour le groupe A et NR pour le groupe B) est celle qui profite le plus du passage au modèle multivarié, puisqu'elle s'aligne sur la précision de l'autre variable, qui est elle-même améliorée (Figure 12).

Le modèle G-BLUP multivarié est donc le meilleur, nous le retenons pour estimer les AGC des 271 géniteurs de l'expérimentation génétique. Sa précision de sélection moyenne observée varie de 0.808 (Groupe B, variable NR) à 0.867 (Groupe A, variables PM et NR).

Remarquons que la corrélation entre les AGC prédites par G-BLUP multivarié et celles prédites par le modèle univarié basé sur le pédigrée varie entre 0.905 (Groupe A, variable PM) et 0.969 (Groupe B, variable PM).

De la même façon, la corrélation entre les ACG prédites par G-BLUP multivarié et celles prédites par le G-BLUP univarié varie entre 0.946 (Groupe B, variable NR) et 0.980 (Groupe A, variable NR).

3.1.3. Etude des résidus

Les postulats sur les résidus du modèle G-BLUP multivarié complet sont vérifiés en nous appuyant sur les graphiques de l'Annexe A. La distribution des résidus de NR s'approche fortement d'une distribution normale, alors que les queues de distribution des résidus de PM s'en écartent (Cf Annexe A, QQ-Plots). Les deux distributions sont quand même unimodales et symétriques (Cf Annexe A, Histogrammes). Pour les deux variables, les résidus sont bien centrés sur zéro, homoscédastiques et indépendants (Cf Annexe A, Résidus en fonction des prédictions).

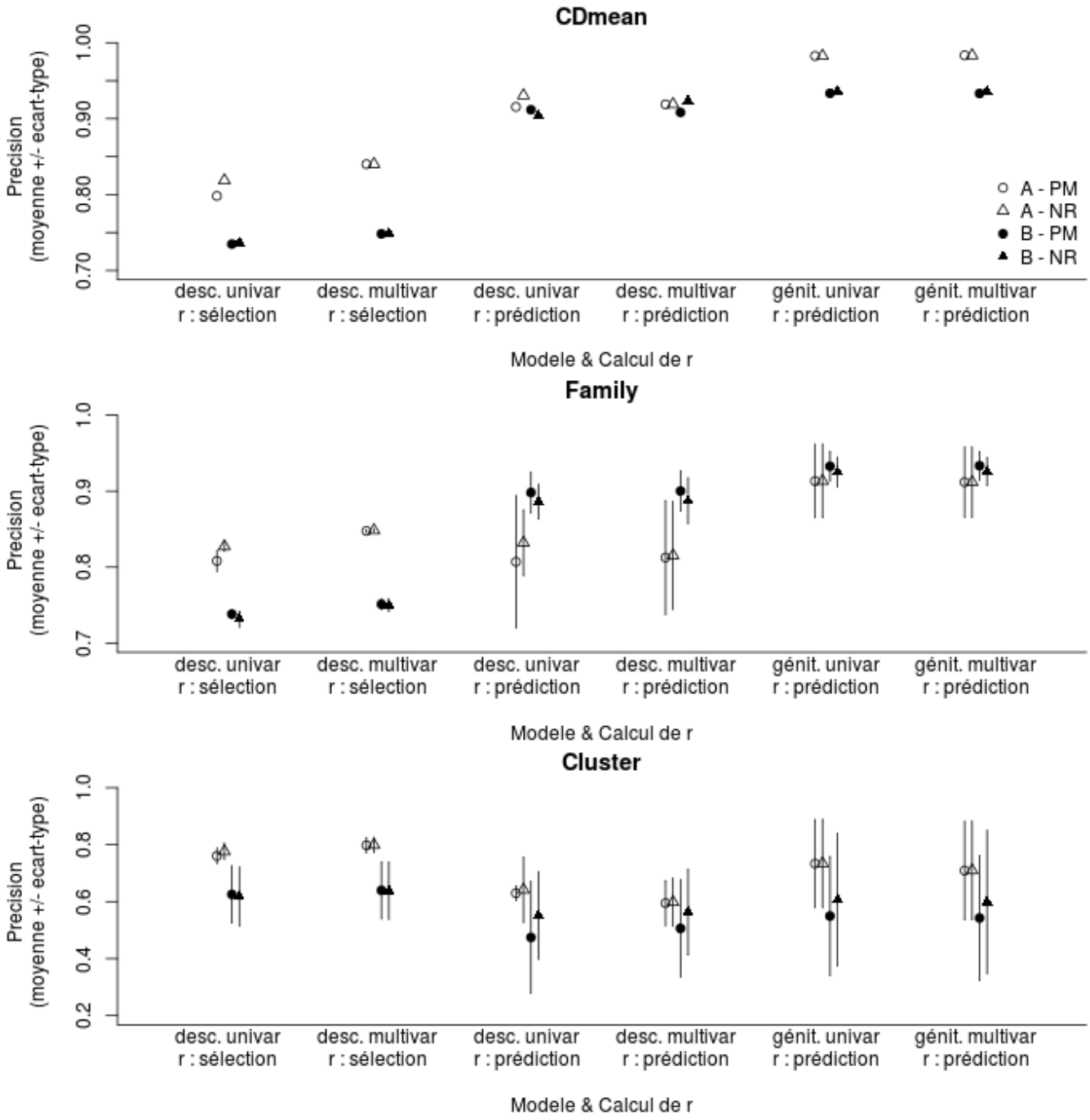


Figure 14 : Précision (r) en fonction du modèle utilisé : sur descendance (desc.) ou sur géniteurs (génit.), univariés (univar.) ou multivarié (multivar.) ; et pour chacune des techniques de VC. Les précisions sont calculées pour ceux des 131 géniteurs Deli du groupe A ou des 131 géniteurs du groupe B qui composent le jeu de validation. Pour le modèle sur descendance, la précision de sélection et la précision de prédiction sont présentées. Pour le modèle sur géniteurs, seule la précision de prédiction est présentée.

3.2. Optimiser le modèle de SG pour des individus non-testés en croisement (VC)

3.2.1. Choix de la matrice d'apparentement

Pour la VC du modèle sur descendance, nous utilisons la matrice $\mathbf{G_{EM}}$ sur la base des résultats précédents (Cf 3.1.1. Choix de la matrice d'apparentement).

Pour la VC du modèle sur géniteurs, nous nous servons de la précision de prédiction du modèle univarié pour identifier laquelle des matrices d'apparentement ($\mathbf{G_{EM}}$, $\mathbf{G_{OF}}$, $\mathbf{G_N}$, avec ou sans l'application de la fonction nearPD) est la meilleure, en utilisant le modèle basé sur le pédigrée (**A**) comme témoin (Figure 13).

La fonction nearPD n'affecte ni le modèle basé sur l'apparentement généalogique (**A**) ni le modèle basé sur $\mathbf{G_{EM}}$. Les modèles basés sur les matrices de VanRaden ($\mathbf{G_{OF}}$, $\mathbf{G_N}$) sont améliorés par l'emploi de cette fonction : augmentation du taux de convergence, amélioration de la précision de prédiction moyenne. Les précisions de prédiction obtenues avec $\mathbf{G_{EM}}$ et avec les matrices de VanRaden auxquelles est appliquée la fonction nearPD sont strictement identiques entre elles, et en moyenne supérieures aux précisions de prédiction obtenues avec **A**, comme nous pouvons le voir sur la Figure 13. Cependant le modèle basé sur **A** converge dans 63.6% des cas alors que les modèles basés sur $\mathbf{G_{EM}}$ ou sur les matrices de VanRaden auxquelles sont appliquées la fonction nearPD convergent chacun dans 13.6% des cas uniquement.

Nous préférons $\mathbf{G_{EM}}$ aux matrices de VanRaden parce que $\mathbf{G_{EM}}$ ne dépend pas de la fonction nearPD pour assurer ses performances. Nous n'utiliserons donc jamais cette fonction. Enfin, nous préférons $\mathbf{G_{EM}}$ à **A** parce que, malgré le faible taux de convergence du modèle basé sur la matrice génomique, il permet de meilleures précisions de prédiction. Les données moléculaires permettent donc de prédire l'AGC d'individus non-testés en croisement plus précisément que le modèle témoin utilisant uniquement le pédigrée. Nous utiliserons par la suite la matrice $\mathbf{G_{EM}}$ pour les modèles G-BLUP sur géniteurs.

3.2.2. Effet du jeu de validation

Les précisions de sélection et de prédiction en VC des modèles G-BLUP univariés et multivarié sur descendance et les précisions de prédiction des modèles G-BLUP univariés et multivarié sur géniteurs sont présentées en Figure 14.

Pour un modèle, un mode de calcul de la précision et une technique de VC donnés, Les précisions des variables PM et NR au sein d'un même groupe (Deli ou B) sont assez similaires.

Aussi bien pour les modèles sur descendance que sur géniteurs, les techniques de VC qui maximisent l'apparentement entre les géniteurs des jeux d'apprentissage et de validation (CDmean, family) augmentent la précision de prédiction, et inversement la technique de VC qui minimise l'apparentement entre les géniteurs des deux jeux (cluster) diminue la précision de prédiction.

En revanche, la précision de sélection est plus stable face à la technique de VC. Elle reste ~0.8 pour le groupe Deli et ~0.7 à ~0.6 pour le groupe B, avec des variations moins marquées que celles que l'on observe pour la précision de prédiction lorsque l'on change la technique de VC. De plus, les barres d'erreur sont plus resserrées pour la précision de sélection, témoignant d'une plus faible dispersion que n'est dispersée la précision de prédiction.

3.2.3. Mode de calcul de la précision

Le gain de précision de sélection que nous avons observé lors du passage au modèle complet univarié au modèle complet multivarié est encore observé en VC (Figure 14). En effet pour toute combinaison de technique de VC, de groupe (Deli ou B), de variable (PM ou NR), la précision de sélection du modèle sur descendance multivarié est meilleure que la précision de sélection du modèle sur descendance univarié.

Cependant, lorsque l'on compare la précision de prédiction des modèles univariés et multivariés, on n'observe pas de différence, que ce soit avec le modèle sur descendance ou avec le modèle sur géniteurs.

3.2.4. Comparaison des modèles sur descendance et sur géniteurs

Il apparaît que le modèle sur géniteurs apporte une meilleure précision de prédiction moyenne que le modèle sur descendance, et ce pour toute combinaison de groupe (A ou B), de variable (PM ou NR), de type de VC, de modèle (univarié ou multivarié) (Figure 14). C'est avec le groupe A que cette supériorité du modèle sur géniteurs par rapport au modèle sur descendance est la plus marquée.

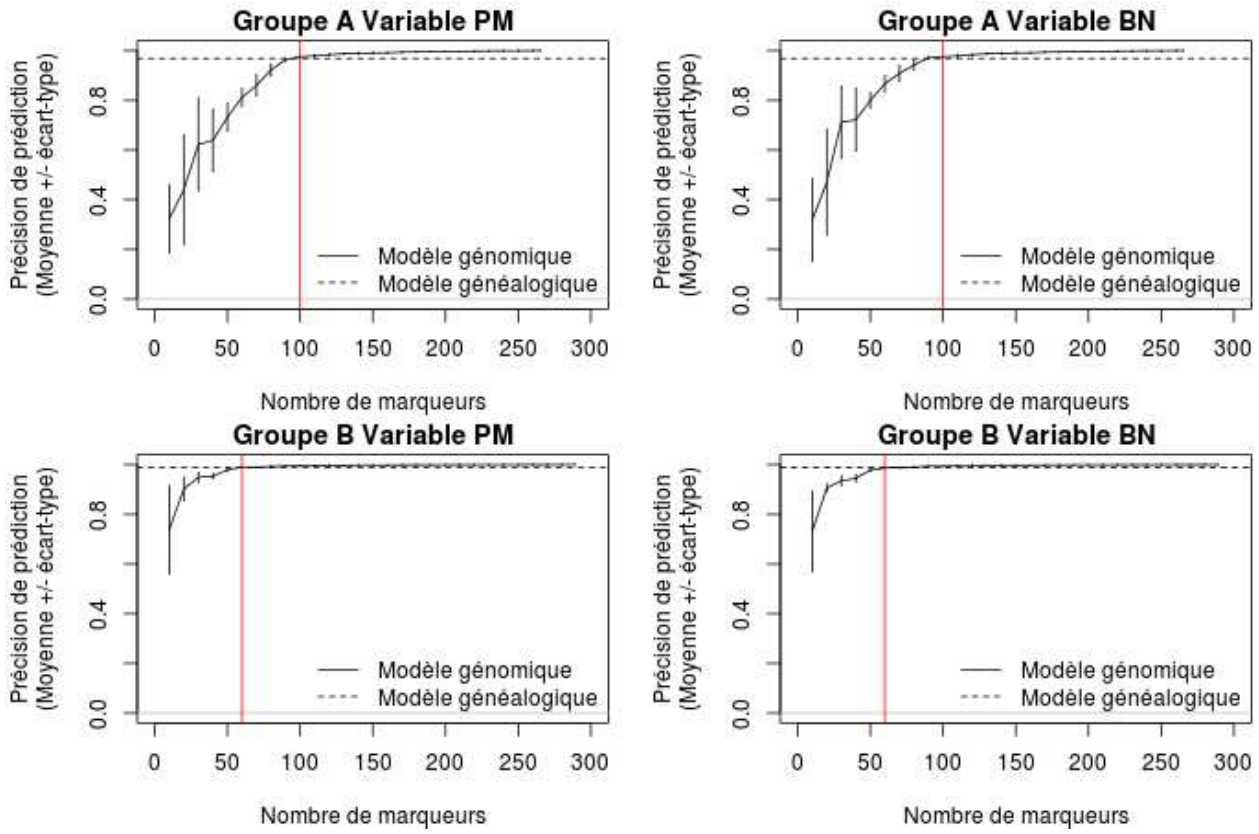


Figure 15 : Evolution de la précision de prédiction du modèle complet sur descendance en fonction du nombre de marqueurs. L'écart-type est calculé sur les 5 répétitions de tirage au sort pour chaque densité de marqueurs. En rouge, le nombre de marqueurs à partir duquel le modèle génomique et le modèle généalogique ont la même précision de prédiction.

Tableau 5, 6, 7 et 8 : Résultats des tests HSD de Tukey pour les précisions de prédiction moyennes des 11 itérations de VC des modèles sur géniteurs : BLUP univarié utilisant le pedigree, GBLUP univarié, GBLUP multivarié et BayesB univarié, et sous différentes densités de marqueurs moléculaires.

Groupe A - PM	10 marq.	20 marq.	30 marq.	40 marq.	70 marq.	80 marq	90 marq.	100 marq.	110 marq.	120 marq.	130 marq.	265 marq.
Pedigree	0.71 (a)	0.71 (a)	0.71 (a)	0.71 (a)	0.71 (a)	0.71 (a)	0.71 (a)	0.71 (a)	0.71 (b)	0.71 (a)	0.71 (b)	0.71 (b)
GBLUP_univar	-0.09 (c)	-0.16 (b)	-0.22 (b)	-0.29 (c)	-0.09 (c)	0.16 (b)	0.64 (b)	0.74 (a)	0.77 (a)	0.74 (a)	0.78 (a)	0.83 (a)
GBLUP_multivar	-0.09 (c)	-0.09 (b)	-0.02 (b)	0.13 (b)	0.50 (b)	0.60 (a)	0.68 (ab)	0.73 (a)	0.75 (ab)	0.73 (a)	0.77 (a)	0.82 (a)
BayesB	0.43 (b)	0.55 (a)	0.55 (a)	0.62 (a)	0.69 (a)	0.67 (a)	0.72 (a)	0.74 (a)	0.77 (a)	0.74 (a)	0.76 (a)	0.80 (a)

Groupe A - NR	10 marq.	20 marq.	30 marq.	40 marq.	70 marq.	80 marq	90 marq.	100 marq.	110 marq.	120 marq.	130 marq.	265 marq.
Pedigree	0.71 (a)	0.71 (a)	0.71 (a)	0.71 (a)	0.71 (a)	0.71 (a)	0.71 (a)	0.71 (a)	0.71 (b)	0.71 (a)	0.71 (b)	0.71 (b)
GBLUP_univar	-0.09 (c)	-0.16 (b)	-0.22 (b)	-0.29 (c)	-0.09 (c)	0.16 (b)	0.64 (b)	0.74 (a)	0.77 (a)	0.74 (a)	0.78 (a)	0.83 (a)
GBLUP_multivar	-0.10 (c)	-0.08 (b)	-0.03 (b)	0.13 (b)	0.50 (b)	0.60 (a)	0.68 (ab)	0.73 (a)	0.75 (ab)	0.73 (a)	0.77 (a)	0.82 (a)
BayesB	0.43 (b)	0.56 (a)	0.55 (a)	0.62 (a)	0.69 (a)	0.67 (a)	0.72 (a)	0.73 (a)	0.77 (a)	0.74 (a)	0.76 (a)	0.80 (a)

Groupe B - PM	10 marq.	20 marq.	30 marq.	40 marq.	50 marq.	60 marq.	289 marq.
Pedigree	0.42 (a)	0.42 (ab)	0.42 (ab)	0.42 (a)	0.42 (a)	0.42 (b)	0.42 (b)
GBLUP_univar	-0.06 (b)	0.02 (c)	0.30 (b)	0.43 (a)	0.49 (a)	0.65 (a)	0.76 (a)
GBLUP_multivar	-0.03 (b)	0.22 (bc)	0.33 (b)	0.44 (a)	0.52 (a)	0.65 (a)	0.76 (a)
BayesB	0.57 (a)	0.59 (a)	0.64 (a)	0.64 (a)	0.61 (a)	0.67 (a)	0.75 (a)

Groupe B - NR	10 marq.	20 marq.	30 marq.	40 marq.	50 marq.	60 marq.	289 marq.
Pedigree	0.40 (a)	0.40 (ab)	0.40 (ab)	0.40 (b)	0.40 (b)	0.40 (b)	0.40 (b)
GBLUP_univar	-0.06 (b)	-0.02 (c)	0.24 (b)	0.38 (b)	0.49 (ab)	0.67 (a)	0.78 (a)
GBLUP_multivar	0.04 (b)	0.20 (bc)	0.32 (b)	0.43 (ab)	0.53 (ab)	0.68 (a)	0.78 (a)
BayesB	0.59 (a)	0.62 (a)	0.67 (a)	0.65 (a)	0.64 (a)	0.69 (a)	0.77 (a)

Pour chaque colonne, les précisions moyennes de chacun des modèles sont significativement différentes si elles ne partagent pas de lettre (a), (b) ou (c). La lettre (a) est attribuée à la meilleure moyenne, la lettre (c) à la moins bonne moyenne. Les couleurs sont associées aux lettres (vert : (a), jaune : (b), orange : (c)). Le risque α est fixé à 5%.

3.3. Diminution du nombre de marqueurs

3.3.1. Effet du nombre de marqueurs sur la précision du modèle complet sur descendance

La Figure 15 présente l'évolution de la précision de prédiction du meilleur modèle complet sur descendance, *i.e.* génomique multivarié, lorsque varie le nombre de marqueurs dans chacun des groupes, et pour chacune des variables. Pour chaque répétition de tirage au sort, la variabilité de la précision est traduite par l'écart-type sur les 5 répétitions. Nous faisons également apparaître la précision de prédiction du modèle généalogique multivarié complet sur descendance pour la comparaison. La précision de prédiction du modèle généalogique multivarié est très proche de 1. Elle varie de 0.971 pour les NR et PM du groupe A à 0.986 pour le PM du groupe B.

PM et NR étant très corrélés, les précisions sont presque identiques pour ces deux variables et pour un groupe donné. Nous remarquons que dans les deux cas, la précision croît rapidement et avec une forte dispersion, jusqu'à atteindre un pallier auquel elle se stabilise. Après ce palier, la précision de prédiction du modèle génomique devient équivalente à la précision de prédiction du modèle généalogique. Naturellement, avec une densité de marqueurs suffisamment forte, la précision de prédiction du modèle génomique égale 1 et son écart-type tend vers 0 (la précision de prédiction étant calculée comme une corrélation des AGC avec les AGC calculées en utilisant l'ensemble des marqueurs).

Le groupe A atteint son pallier aux alentours de 60 marqueurs, et le groupe B atteint son pallier aux alentours de 100 marqueurs.

3.3.2. Effet du nombre de marqueurs sur la précision du modèle sur géniteurs en validation croisée

Les précisions de prédiction des 5 répétitions de tirage au sort de marqueurs ont été moyennées à chaque densité, et pour chaque combinaison de groupe (A ou B), de variable (PM ou NR), de jeu de VC et de modèle. Ces précisions moyennes sont ensuite comparées selon le modèle génomique (G-BLUP univarié, G-BLUP multivarié, BayesB univarié) ou témoin (pédigrée univarié) dont elles sont issues par un test HSD de Tukey (p-value<0.05, $n=11$ jeux de VC). Les résultats sont présentés dans les Tableaux 5 à 8. Le détail de l'évolution de la précision de prédiction des 4 modèles est présenté en Annexe C.

Du fait de la forte corrélation entre PM et NR, les précisions moyennes et les résultats des tests sont presque similaires pour les deux variables au sein d'un groupe donné.

Pour les géniteurs du groupe A (Tableaux 5 et 6), on note qu'à partir de 130 marqueurs tous les modèles génomiques sont significativement plus précis que le modèle généalogique. Aucun modèle génomique ne se démarque des autres au dessus de 130 marqueurs. Avec moins de 130 marqueurs les modèles génomiques peuvent être équivalents ou significativement moins précis que le modèle généalogique. Pour les faibles densités de marqueurs, le modèle BayesB est le meilleur modèle génomique. De 10 à 70 marqueurs, il s'avère significativement plus précis que les modèles G-BLUP univarié et multivariés. Il n'y a qu'à la densité de 10 marqueurs que le modèle généalogique surpasse significativement en précision le modèle BayesB.

Enfin, le modèle G-BLUP multivarié est plus précis que le modèle G-BLUP univarié pour les faibles densités de marqueurs moléculaires. Cette différence apparaît clairement sur les graphiques (Cf Annexe C) et est significative de 40 à 80 marqueurs.

Pour les géniteurs du groupe B (Tableaux 7 et 8), c'est à partir de 60 marqueurs que tous les modèles génomiques se valent et surpassent significativement la précision du modèle généalogique. La précision du modèle BayesB n'est jamais significativement inférieure à la précision du modèle généalogique même lorsque l'on descend à 10 marqueurs. Le modèle BayesB est le modèle génomique le plus précis lorsque l'on descend sous les 60 marqueurs. La supériorité du modèle BayesB par rapport au G-BLUP est significative de 10 à 30 marqueurs moléculaires. A faible densité de marqueurs, le modèle G-BLUP multivarié est plus précis que le modèle G-BLUP univarié mais cette différence n'est pas significative (Cf Annexe C).

Tableau 9 : Héritabilité des caractères PM et NR pour les populations A et B, d'après Cros et al. *(2014).*

h^2	A	B
PM	0.225	0.566
NR	0.306	0.499

4. Discussion

Les études sur la sélection génomique multivariée appliquée à des données empiriques sont peu fréquentes dans la littérature.

Wong et Bernardo (2008), étudiant la réponse à la sélection génomique univariée appliqué au palmier à huile grâce à des simulations, identifient comme étant des facteurs déterminants la taille de la population (n) et l'héritabilité (h^2) des caractères. La supériorité de la SG s'affirme dès n = 50 avec leurs données simulées. Avec nos plus de 130 palmiers dans chaque groupe, nous nous assurons que la taille de la population n'est pas limitante. L'héritabilité des caractères PM et NR pour chacun des groupes est présentée en Tableau 9.

4.1. Lien avec l'héritabilité

D'après les résultats obtenus par Wong et Bernardo (2008), la supériorité de la sélection génomique sur la sélection phénotypique s'accroît lorsque l'héritabilité augmente. Nos résultats diffèrent, avec une précision de sélection génomique plus élevée chez le groupe A que chez le groupe B (pour PM comme pour NR), alors que les héritabilités sont plus élevées dans le groupe B que dans le groupe A (pour PM comme pour NR). Les précisions de prédictions sont aussi, en moyenne, plus élevées dans le groupe A que dans le groupe B. Ceci traduit le fait que d'autres facteurs affectent la précision de la SG, et notamment le nombre de QTL et le déséquilibre de liaison entre marqueurs et QTL, inconnus.

De plus, dans notre étude, bien que les précisions soient toujours plus élevées dans le groupe A que dans le groupe B pour les caractères PM et NR (alors que ces caractères sont plus héritables dans le groupe B que dans le groupe A), on montre que les précisions de prédiction du modèle sur descendance en validation croisée profitent davantage du passage à la sélection génomique dans le groupe B que dans le groupe A.

4.2. Matrices d'apparentement

Que l'on cherche à prédire la valeur génétique de géniteurs testés en croisement ou à étendre les prédictions à des géniteurs non-testés en croisement, le modèle G-BLUP (utilisant la matrice $\mathbf{G_{EM}}$) se distingue positivement du modèle BLUP classique. Cela confirme la supériorité de la sélection génomique par rapport à la sélection classique.

L'extension du calcul de la matrice de VanRaden au cas multi-allélique (comme dans le cas des marqueurs microsatellites) grâce à la technique d'Andrés Legarra est inédite. Cependant, l'impossibilité de fournir à cette matrice une interprétation génétique (coefficients d'apparentement négatifs pour la moitié) et l'inadéquation à notre jeu de données (performances statistiques inférieures à celles obtenues avec les autres matrices) nous ont forcés à l'écarter. De toutes les matrices testées, la meilleure matrice d'apparentement génomique avec des données réelles de microsatellites est celle proposée par Eding et Meuwissen (2001) $\mathbf{G_{EM}}$.

La matrice $\mathbf{G_{EM}}$ contient des valeurs d'apparentement plus élevées que la matrice **A**. D'un côté, cela signifie sûrement que l'utilisation des marqueurs permet de tenir compte d'une généalogie plus profonde que celle apparente dans le pédigrée. L'estimateur d'apparentement génomique révèle des apparentements et des consanguinités ignorées par le pédigrée, d'où la meilleure performance statistique du modèle génomique complet sur descendance par rapport au modèle généalogique. D'un autre côté, nous avons fait l'hypothèse qu'aucun allèle n'était AIS lors du calcul de $\mathbf{G_{EM}}$. C'est une hypothèse irréalisable en pratique, qui provoque une surestimation des coefficients d'apparentement.

L'AIS (*alikeness in state*) peut réduire l'efficacité de l'estimateur d'Eding et Meuwissen. En effet les allèles AIS peuvent être apparus par mutation, donc sans qu'un ancêtre commun ne les ait transmis, ou être hérités d'un ancêtre très éloigné, de sorte qu'ils ne soient plus en DL avec les mêmes allèles des QTL. Maenhout *et al.* (2009) ont développé une matrice $\mathbf{G_{WAIS}}$, pour « *weighted alikeness in state* », permettant de prendre en compte l'AIS. Ils garantissent que, comme $\mathbf{G_{EM}}$ (que l'on appellerait $\mathbf{G_{AIS}}$ d'après leur article), la matrice $\mathbf{G_{WAIS}}$ est semi-définie positive, et que ses coefficients d'apparentement sont bien compris entre 0 et 2 (les coefficients de parentés de Malécot associés peuvent être interprétés comme des probabilités). Pour approfondir l'étude, il serait intéressant de comparer les performances des matrices $\mathbf{G_{AIS}}$ *vs.* $\mathbf{G_{WAIS}}$, en termes de vraisemblance et de précision des modèles G-BLUP qui les utiliseraient.

Notre modèle génomique aurait également pu être amélioré par l'utilisation d'une matrice de dominance **D** calculée à partir des marqueurs moléculaires. Par exemple, Zapata-Valenzuela *et al.* (2012) utilisent un modèle linéaire mixte pour calculer les effets associés aux marqueurs, en considérant non seulement les effets additifs au marqueur mais aussi les effets de dominance au marqueur.

4.3. Modèles multivariés

Le passage au modèle multivarié montre des résultats contradictoires. La précision de sélection du modèle complet sur descendance calculée par la formule utilisant la VEP atteste un gain élevé et significatif lors du passage au multivarié. Cependant (i) les corrélations entre les AGC des palmiers prédites par les modèles complets sur descendance univariés et multivarié sont très proches de 1 ; (ii) de fait les précisions de prédictions calculées comme des corrélations entre les AGC prédites en univariés et en multivarié ne sont pas affectées par le passage au modèle multivarié ; (iii) et enfin, en changeant de technique de validation croisée la précision de sélection reste relativement stable alors que la précision de prédiction varie beaucoup.

Le meilleur moyen de mettre à l'évidence – ou non – le gain de précision de la sélection génomique lors du passage au modèle multivarié serait de réaliser des simulations. La précision de sélection serait ainsi calculée comme une corrélation entre TBV et GEBV, sans ambigüité. Les résultats de simulations disponibles dans la littérature tendent à prouver un gain de précision en multivarié (Calus et Veerkamp, 2011; Jia et Jannink, 2012; Guo *et al.*, 2014), allant donc dans le sens des résultats de précision de sélection obtenus par la VEP. Peut-être la précision de sélection calculée à partir de la VEP doit-elle être interprétée comme une mesure de fiabilité des prédictions des AGC, sans pour autant bouleverser les estimations de la valeur génétique additive des palmiers.

Au moyen de simulations, Guo *et al.* (2014) concluent qu'à partir d'une héritabilité $h^2 = 0.3$, le modèle G-BLUP multivarié n'apporte pas plus de précision de sélection que le modèle G-BLUP univarié. En revanche, le gain de précision est mis à l'évidence dans le cas où l'un des caractères présente 90% de valeurs manquantes, ou alors pour un caractère d'une très faible héritabilité ($h^2 = 0.05$). La forte héritabilité de nos caractères d'intérêt, PM et NR (Cf Tableau 9) pourrait expliquer la faible amplitude du gain de précision de sélection lors du passage au modèle multivarié.

Conformément à la littérature, nous remarquons que c'est au sein de chaque groupe la variable la moins héritable (PM pour le groupe A et NR pour le groupe B) qui profite le plus du passage au modèle multivarié lorsque l'on compare les précisions de sélection, calculées avec la VEP, du modèle complet sur descendance.

Remarquons enfin qu'à faible densité de marqueurs, le modèle G-BLUP multivarié est plus performant que le modèle G-BLUP univarié. Cependant dans ces conditions le modèle BayesB univarié est encore meilleur. Il pourrait être intéressant de s'interroger sur un modèle Bayésien multivarié. En effet, le modèle G-BLUP est limité par une hypothèse forte sur laquelle il se base : la similarité de la distribution de tous les effets aux marqueurs. De fait dans le modèle multivarié la covariance entre les caractères est considérée identique à chaque marqueur. Cela pourrait limiter l'intérêt du modèle G-BLUP multivarié (Guo *et al.*, 2014). Dans le cas d'une architecture génétique dans laquelle des QTL majeurs interviennent, le passage au multivarié est beaucoup plus avantageux pour les modèles Bayésiens que pour le modèle G-BLUP (Jia et Jannink, 2012).

4.4. Prédiction de l’AGC de géniteurs non-testés en croisement

L’utilisation de stratégies de définition des jeux de validations croisées contrastées permet de donner un aperçu de ce qu’il faut attendre des prédictions d’AGC de géniteurs non-testés en croisement. L’interprétation de la précision de sélection calculée par la formule de la VEP étant délicate, nous nous intéressons surtout aux résultats obtenus avec la précision de prédiction, calculée par corrélation avec les AGC de référence.

Ces résultats montrent la supériorité du modèle sur géniteur par rapport au modèle sur descendance. Cela est d’autant plus intéressant que le modèle sur géniteur est plus simple à mettre en œuvre, notamment en termes de temps de calcul. Nous supposons que l’optimisation du plan de croisement qui a été réalisée lors de la mise en place du test génétique rend le modèle sur descendance sensible à la soustraction d’un jeu de validation, ce qui expliquerait la faible robustesse de ce modèle en validation croisée.

Nous mettons aussi en avant la nécessité de maximiser l’apparentement entre le jeu d’apprentissage et le jeu de validation afin d’obtenir une bonne précision de prédiction. Cela signifie que pour prédire les AGC de géniteurs qui n’auraient pas été testés en croisement à partir des AGC des géniteurs impliqués dans un test génétique, il convient de s’assurer que les nouveaux géniteurs soient le plus apparentés possibles aux géniteurs du test génétique, ou dans le cas contraire, de rester critique sur ces extrapolations.

Notons que le modèle sur géniteurs n’a pas été pondéré. En d’autres termes, un géniteur représenté par n_1 descendants et un géniteur représenté par n_2 descendants sont considérés comme étant aussi informatifs dans le modèle sur géniteurs, même si $n_1 \neq n_2$. Les études précédentes mobilisant ce jeu de données utilisaient une pondération individuelle tenant compte de la fiabilité de l’AGC (Souchard, 2013; Cros *et al.*, 2014). Cependant, nous avons vérifié que les variations dans le nombre de descendants entre géniteurs n’étaient pas suffisamment importantes pour vraiment affecter la fiabilité des AGC.

Garrick *et al.* (2009) proposent un argumentaire en faveur de la dérégression. La dérégression est une technique utilisée pour retirer les effets parentaux moyens, dans le cas justement d’un modèle dont les données d’entrée sont des valeurs génétiques additives (des EBV ou des AGC). La dérégression est utilisée par Cros *et al*. (2014), mais dans le cas de notre jeu de données, elle a un effet très faible sur les AGC (probablement car leur calcul repose sur un grand nombre d’observations phénotypiques sur les descendants) et n’a donc pas été employée dans cette étude.

4.5. Nombre de marqueurs

Tout comme Zapata-Valenzuela *et al.* (2012), nous constatons qu'en utilisant seulement un sous-échantillon de marqueurs la qualité du modèle génomique n'est pas affectée, à condition que ce sous-échantillon ait une taille minimale. Zapata-Valenzuela *et al.* estiment un effet au marqueur, et obtiennent les mêmes résultats en constituant leur sous-échantillon avec des marqueurs dont les effets sont significatifs, ou avec des marqueurs sélectionnés aléatoirement. Ils expliquent ce résultat en indiquant que le sous-échantillon de marqueurs, qu'ils soient en DL avec les QTL ou non, reconstruit implicitement la matrice d'apparentement **G** entre les géniteurs. Dans le cas du G-BLUP, la construction de **G** est explicite. Il serait tout de même intéressant de confirmer cette hypothèse en comparant la performance de sous-échantillons de marqueurs sélectionnés aléatoirement à la performance de sous-échantillons de marqueurs en DL avec les QTL, ou dont l'effet associé est significatif.

La taille minimale de l'échantillon de marqueurs nécessaire varie selon la population. Elle est à peu près identique lorsque l'on enquête sur le modèle complet sur descendance ou sur le modèle sur géniteurs en validation croisée. Pour le groupe A, la qualité du modèle est significativement affectée lorsque l'on utilise moins de 130 marqueurs, et pour le groupe B, lorsque l'on descend sous les 60 marqueurs. L'origine de cette différence est certainement à rechercher dans l'histoire des deux groupes. Le groupe A comprend essentiellement des Deli (131 sur 140 palmiers) qui sont issus de 4 palmiers fondateurs et forment par conséquent une population très consanguine. Cela explique qu'il soit nécessaire de mobiliser plus de marqueurs pour distinguer les relations d'apparentement. Au contraire, le groupe B a une base génétique plus large car il est constitué d'un plus grand nombre de populations, mieux représentées.

La possibilité d'utiliser moins de marqueurs pour une performance équivalente présente un intérêt économique indéniable.

Conclusion

Ce mémoire illustre une fois de plus l'intérêt de la sélection génomique. Le modèle G-BLUP basé sur l'estimateur d'apparentement d'Eding et Meuwissen (2001) s'avère plus performant que le modèle mixte traditionnel basé sur le pédigrée. Le modèle G-BLUP est à la fois plus performant pour analyser un test génétique, et pour prédire la valeur génétique additive de palmiers non-testés en croisement.

Nous montrons que la précision de sélection calculée à partir de la VEP augmente grâce au modèle multivarié. Cependant la corrélation entre les AGC prédites par les modèles univariés et multivariés ne traduit pas une amélioration aussi élevée que celle attendue grâce à la formule de la VEP. La forte héritabilité des caractères PM et NR pourrait être en cause. Afin d'approfondir notre étude, il pourrait être intéressant de réaliser des simulations afin de calculer avec exactitude la précision de sélection. Il faudrait aussi enquêter sur le modèle multivarié Bayésien.

Enfin, nous montrons que 130 marqueurs pour le groupe A et 60 marqueurs sont suffisants pour assurer les performances du modèle G-BLUP, que celui-ci soit utilisé pour un test génétique ou pour prédire la valeur génétique additive de palmiers non-testés en croisement.

Références bibliographiques

I. Livres

Falconer D. et Mackay T., 1996. ***Introduction to quantitative genetics***. Longman, Harlow, Essex, UK, 464 p.

Gallais A., 1990. ***Théorie de la sélection en amélioration des plantes***. Masson.

Gilmour A.R., Gogel B.J., Cullis B.R. et Thompson R., 2009. ***ASReml user guide release 3.0***. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK www.vsni.co.uk

Mrode R.A., 2005. ***Linear models for the prediction of animal breeding values***. CABI, Oxfordshire, UK, 344 p.

II. Articles

Bernardo R., 1996. **Best linear unbiased prediction of maize single-cross performance**. Crop Science, 36(1): 50‑56.

Billotte N., Marseillac N., Risterucci A.-M., Adon B., Brottier P. *et al.*, 2005. **Microsatellite-based high density linkage map in oil palm (*Elaeis guineensis* Jacq.).** Theoretical and Applied Genetics, 110(4): 754‑765.

Browning S.R. et Browning B.L., 2007. **Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering**. The American Journal of Human Genetics, 81(5): 1084‑1097.

Calus M. et Veerkamp R., 2011. **Accuracy of multi-trait genomic selection using different methods**. Genetics Selection Evolution, 43(1): 26.

De los Campos G., Hickey J.M., Pong-Wong R., Daetwyler H.D. et Calus M.P.L., 2013. **Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding**. Genetics, 193(2): 327‑345.

Clark S.A., Hickey J.M., Daetwyler H.D. et van der Werf J.H., 2012. **The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes**. Genet Sel Evol, 44(4):

Cros D., Denis M., Sánchez L., Cochard B., Flori A. *et al.*, 2014. **Genomic selection prediction accuracy in a perennial crop: case study of oil palm (Elaeis guineensis Jacq.)**. Under review.

Daetwyler H.D., Calus M.P.L., Pong-Wong R., de los Campos G. et Hickey J.M., 2013. **Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking**. Genetics, 193(2): 347‑365.

Durand-Gasselin T., Kouame Kouame R., Cochard B., Adon B. et Amblard P., 2000. **Diffusion variétale du palmier à huile (*Elaeis guineensis* Jacq.)**. Oléagineux, Corps Gras, Lipides, 7(2): 207‑214.

Eding H. et Meuwissen T.H.E., 2001. **Marker-based estimates of between and within population kinships for the conservation of genetic diversity**. Journal of Animal Breeding and Genetics, 118(3): 141‑159.

Forni S., Aguilar I. et Misztal I., 2011. **Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information**. Genetics Selection Evolution, 43(1): 1.

Garrick D.J., Taylor J.F. et Fernando R.L., 2009. **Deregressing estimated breeding values and weighting information for genomic regression analyses**. Genet Sel Evol, 41(55): 44.

Gascon J. et de Berchoux C., 1964. **Caractéristique de la production d'Elaeis guineensis (Jacq.) de diverses origines et de leurs croisements. Application à la sélection du palmier à huile.** Oléagineux, 19(2): 75‑84.

Guo G., Zhao F., Wang Y., Zhang Y., Du L. et Su G., 2014. **Comparison of single-trait and multiple-trait genomic prediction models**. BMC genetics, 15(1): 30.

Hayes B.J., Bowman P.J., Chamberlain A.J. et Goddard M.E., 2009. **Invited review: Genomic selection in dairy cattle: Progress and challenges**. Journal of Dairy Science, 92(2): 433‑443.

Heffner E.L., Sorrells M.E. et Jannink J.L., 2009. **Genomic selection for crop improvement**. Crop Sci, 49: 1 ‑ 12.

Henderson C.R., 1975. **Best Linear Unbiased Estimation and Prediction under a Selection Model**. Biometrics, 31: 423‑447.

Jia Y. et Jannink J.-L., 2012. **Multiple-Trait Genomic Selection Methods Increase Genetic Value Prediction Accuracy**. Genetics, 192(4): 1513‑1522.

Lorrenz A.J., Chao S., Asoro F.G., Heffner E.L., Hayashi T., Iwata H., Smith K.P., Sorrells M.E. et Jannink J.-L., 2011. **Genomic Selection in Plant Breeding : Knowledge and Prospects**. Advances in Agronomy, 110: 77‑123.

Maenhout S., De Baets B. et Haesaert G., 2009. **Marker-based estimation of the coefficient of coancestry in hybrid breeding programmes**. Theoretical and applied genetics, 118(6): 1181‑1192.

Meuwissen T.H.E., Hayes B.J. et Goddard M.E., 2001. **Prediction of total genetic value using genome-wide dense marker maps**. Genetics, 157(4): 1819‑1829.

Nakaya A. et Isobe S.N., 2012. **Will genomic selection be a practical method for plant breeding?** Annals of Botany.

Ovaskainen O., Cano J.M. et Merilä J., 2008. **A Bayesian framework for comparative quantitative genetics**. Proceedings of the Royal Society B: Biological Sciences, 275(1635): 669‑678.

Piepho H.P., Möhring J., Melchinger A.E. et Büchse A., 2008. **BLUP for phenotypic selection in plant breeding and variety testing**. Euphytica, 161(1-2): 209‑228.

Rincent R., Laloë D., Nicolas S., Altmann T., Brunel D. *et al.*, 2012. **Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.)**. Genetics.

Stuber C.W. et Cockerham C.C., 1966. **Gene effects and variances in hybrid populations**. Genetics, 54(6): 1279.

Toro M.Á., García-Cortés L.A. et Legarra A., 2011. **A note on the rationale for estimating genealogical coancestry from molecular markers**. Genet Sel Evol, 43(1): 27.

Tranbarger T.J., Kluabmongkol W., Sangsrakru D., Morcillo F., Tregear J.W., Tragoonrung S. et Billotte N., 2012. **SSR markers in transcripts of genes linked to post-transcriptional and transcriptional regulatory functions during vegetative and reproductive development of Elaeis guineensis**. BMC plant biology, 12(1): 1.

VanRaden P.M., 2007. **Genomic measures of relationship and inbreeding**. Interbull Bulletin, 37: 33‑36.

Wimmer V., Albrecht T., Auinger H.-J. et Schön C.-C., 2012. **Synbreed: a framework for the analysis of genomic prediction data using R**. Bioinformatics, 28(15): 2086‑2087.

Wolak M.E., 2012. **Nadiv: an R package to create relatedness matrices for estimating non-additive genetic variances in animal models**. Methods in Ecology and Evolution, 3(5): 792‑796.

Wong C.K. et Bernardo R., 2008. **Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations**. Theoretical and Applied Genetics, 116(6): 815‑824.

Zaki N.M., Singh R., Rosli R. et Ismail I., 2012. **Elaeis oleifera Genomic-SSR Markers: Exploitation in Oil Palm Germplasm Diversity and Cross-Amplification in Arecaceae**. International Journal of Molecular Sciences, 13(4): 4069–4088.

Zapata-Valenzuela J., Isik F., Maltecca C., Wegrzyn J., Neale D., McKeand S. et Whetten R., 2012. **SNP markers trace familial linkages in a cloned population of Pinus taeda—prospects for genomic selection**. Tree Genetics & Genomes, 8(6): 1307‑1318.

III. Littérature grise

Bates D. et Maechler M., 2014. ***Matrix: Sparse and Dense Matrix Classes and Methods***. R package.

CIRAD, 2014.Tout savoir sur le palmier à huile. http://www.cirad.fr/publications-ressources/science-pour-tous/dossiers/palmier-a-huile/les-enjeux.

Cros D., 2013.Oil palm breeding. .

Denis M., 2012.Les méthodes bayésiennes pour la sélection génomique. .

indexmundi, 2014.Palm Oil Production by Country in 1000 MT - Country Rankings. http://www.indexmundi.com/agriculture/?commodity=palm-oil&.

Isik F. et Whetten R., 2011.Workshop - Genomic Selection in Tree Breeding. .

Legarra A., 2014. ***Multiallelic genomic relationship matrices***.

De Mendiburu F., 2012. ***Agricolae: Statistical Procedures for Agricultural Research***. R package.

MPOB, 2011.Welcome to the Malaysian Palm Oil Board // Malaysian Palm Oil Industry // Washington, DC // 1-202-572-9768. http://www.palmoilworld.org/about_malaysian-industry.html.

R Core Team, 2014. ***R: a language and environment for statistical computing***. R Foundation for Statistical Computing.

Rodriguez P.P. et de los Campos G., 2013. **BGLR: A Statistical Package for Whole-Genome Regression**.

Souchard V., 2013. **Sélection du palmier à huile pour la régularité de la production de régimes**. Université de Montpellier 2 - Sciences et Techniques.

Université de Toulouse, 2014. Analyse de variance multivariée. http://wikistat.fr/pdf/st-m-modmixt5-manova.pdf.

Walsh B., 2013.Lecture 28: BLUP and genomic selection, Bruce Walsh lecture notes, Synbreed course, version 11 July 2013, 55p. .

Table des Annexes

Annexe A : Programmes

a. Préparer une matrice A

```
> # Création de la matrice A pour le groupe A
> pedigreeA <- read.table("pedigree_sire_damA_CONSENSUS.csv", sep=";",header=T)
>
> library(synbreed)
> gpPedigreeA<-create.pedigree(ID=pedigreeA[,1],
+                              Par1=pedigreeA[,2],
+                              Par2=pedigreeA[,3],
+                              gener=NULL,
+                              sex=NULL, add.ancestors=F)
> gpPedigreeA<-create.gpData(pedigree=gpPedigreeA)
> AA<- kin(gpPedigreeA, ret=("add"))
> attr(AA,"class")<-NULL
> AA<-0.5*AA
>
> # Mise au format ASReml
> AA.giv <-write.relationshipMatrix(AA, sorting="ASReml",type="ginv")
```

b. Préparer une matrice D

```
> ped_A<- read.table("pedigree_sire_damA_CONSENSUS.csv", sep=";",header=T)
> ped_B<- read.table("pedigree_sire_damB_CONSENSUS.csv", sep=";",header=T)
> names(ped_A) <- c('PAL','M','P')
> names(ped_B) <- c('PAL','M','P')
> full_pedigree <- rbind(ped_A,ped_B)
>
> library(nadiv)
> D <- makeDsim(full_pedigree,
+               N=10000,
+               parallel = T,
+               ncores = getOption("mc.cores", 2L),
+               invertD =F,
+               calcSE = FALSE,
+               returnA = FALSE)$D
> D <- as.matrix(D)
> rownames(D) <- full_pedigree$PAL
> colnames(D) <- full_pedigree$PAL
>
> # Conservons uniquement les croisements contenus
> # dans le fichier de données bunch_DATA
> temp_pedigree <- bunch_DATA[,c('FAMILLE','parent_A','parent_B')]
> temp_pedigree <- unique(temp_pedigree)
> names(temp_pedigree) <- names(full_pedigree)
> full_pedigree <- rbind(full_pedigree, temp_pedigree)
> D <-D[which(rownames(D) %in% temp_pedigree$PAL),
+       which(colnames(D) %in% temp_pedigree$PAL)]
>
> # Mise au format ASReml
> library(synbreed)
> D.giv <- write.relationshipMatrix(D, sorting='ASReml', type='ginv')
```

c. Réaliser le modèle BLUP univarié sur descendance avec ASReml

```
> # Exemple de modèle BLUP univarié sur descendance pour le poids moyen (ABW)
> library(asreml)
> BLUP.ABW.offspring <- asreml(fixed =  ABW ~ 1 + ESSAI + ESSAI:REPET + age,
+                              random = ~  ped(parent_A)
+                                        + ped(parent_B)
+                                        + ESSAI:EXP
+                                        + giv(FAMILLE)
+                                        + giv(FAMILLE):age
+                                        + PALMIER,
+                              ginverse = list(parent_A=  AA.giv,
+                                              parent_B=  AB.giv,
+                                              FAMILLE=   D.giv),
+                              na.method.X='include',
+                              maxiter = 1000,
+                              workspace=64e+06,
+                              data =  bunch_DATA,
+                              trace=F)
```

d. Préparer une matrice H à partir de G_{EM}

```
> # Exemple pour le groupe A
> ## Création de GEM
> GA <-(X%*%t(X))/(2*n_markers_kept)
> GA<-as.matrix(GA)
> rownames(GA)<-colnames(GA)
> GA<-0.5*GA
>
> ## Création de H-1
> AA<-AA[c(sort(colnames(AA)[which(colnames(AA)%in%colnames(GA)==F)]),
+          sort(colnames(AA)[which(colnames(AA)%in%colnames(GA)==T)])),
+        c(sort(colnames(AA)[which(colnames(AA)%in%colnames(GA)==F)]),
+          sort(colnames(AA)[which(colnames(AA)%in%colnames(GA)==T)]))]
> GA<-GA[sort(colnames(GA)[which(colnames(GA)%in%colnames(AA))]),
>        sort(colnames(GA)[which(colnames(GA)%in%colnames(AA))])]
> AA22<-AA[c(sort(colnames(AA)[which(colnames(AA)%in%colnames(GA)==T)])),
>          c(sort(colnames(AA)[which(colnames(AA)%in%colnames(GA)==T)]))]
> MA<-matrix(0,ncol=ncol(AA),nrow=ncol(AA))
> id1<-length(colnames(AA)[which(colnames(AA)%in%colnames(GA)==F)])
> id2<-length(colnames(AA)[which(colnames(AA)%in%colnames(GA)==T)])
> MA[(id1+1):(id1+id2),(id1+1):(id1+id2)]<-ginv(as.matrix(GA))-
+                                          ginv(as.matrix(AA22))
> H.A.inv<-ginv(as.matrix(AA))+MA
> colnames(H.A.inv)<-rownames(H.A.inv)<-colnames(AA)
>
> ## Application de nearPD (désactivé par des '#')
> # H.A.inv <-as.matrix(nearPD(H.A.inv)$mat)
> # colnames(H.A.inv)<-rownames(H.A.inv)<-colnames(AA)
>
> # Mise au format ASReml
> library(synbreed)
> HA.giv <-write.relationshipMatrix(H.A.inv,sorting="ASReml",type="none")
```

e. Réaliser le modèle G-BLUP multivarié sur descendance avec ASReml

```
> # le fichier a1 contient les 17 paramètres initiaux de variance-covariance
> a1 <- read.table("initialisation_genomic_offspring.txt", header=T, quote="\"")
>
> # GBLUP multivarié sur descendance pour
> # le poids moyen (ABW) et le nombre de régimes (BN)
> library(asreml)
> t1<-Sys.time()
> GBLUP.multi.offspring <- asreml(fixed=cbind(ABW,BN) ~ trait
+                                                     + trait:ESSAI
+                                                     + trait:ESSAI:REPET
+                                                     + trait:age ,
+                           random= ~ us(trait, init=c(a1[1:3])):ped(parent_A)
+                                   + us(trait, init=c(a1[4:6])):ped(parent_B)
+                                 + diag(trait, init=c(a1[7:8])):giv(FAMILLE)
+                                 + diag(trait, init=c(a1[9:10])):giv(FAMILLE):age
+                                 + diag(trait, init=c(a1[11:12])):PALMIER
+                                 + diag(trait, init=c(a1[13:14])):ESSAI:EXP,
+                           ginverse= list(parent_A= HA.giv,
+                                          parent_B= HB.giv,
+                                          FAMILLE=  D.giv),
+                           rcov= ~ units:us(trait, init=c(a1[15:17])),
+                           data= bunch_DATA,
+                           na.method.X="include",
+                           workspace=64e+06,
+                           maxiter=1000,
+                           trace=F)
> t2<-Sys.time()
> t2-t1
```

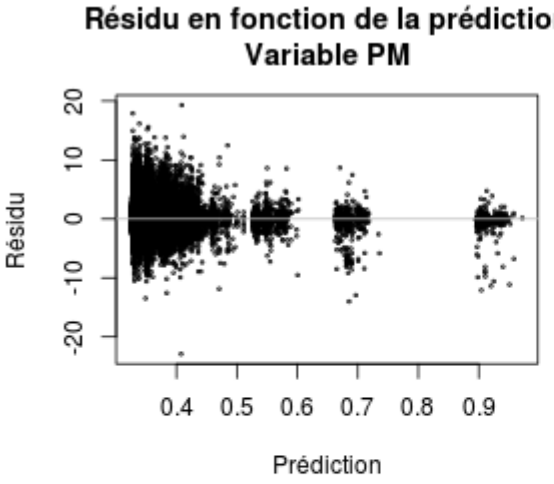
f. Réaliser le modèle G-BLUP multivarié sur géniteurs avec ASReml

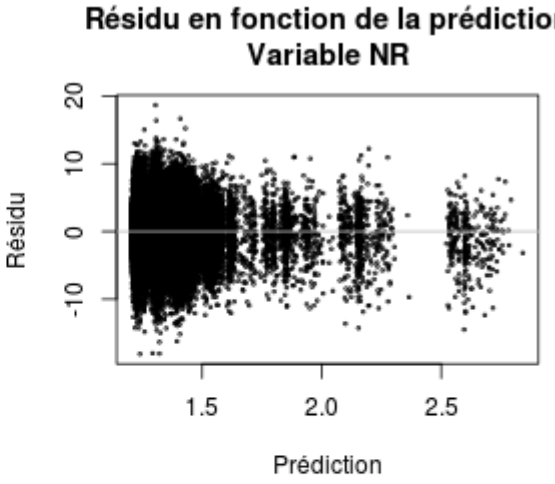
```
> # Préparation d'une matrice contenant uniquement l'apparentement moléculaire
> library(synbreed)
> GA.giv<-write.relationshipMatrix(GA, sorting="ASReml",type="ginv")
>
> # Préparation d'une matrice contenant uniquement l'apparentement moléculaire
> a1 <- read.table("initialisation_genomic_genitors_A.txt", header=T, quote="\"")
>
> # GBLUP multivarié sur géniteurs
> library(asreml)
> GBLUP.multi.genitors <- asreml(fixed = cbind(ABW,BN) ~ trait,
+                          random = ~ us(trait,init=a1[1:3]):giv(animal),
+                          rcov = ~ units:us(trait,init=a1[4:6]),
+                          ginverse=list(animal=GA.giv),
+                          data=DATA,
+                          maxiter=1000,
+                          trace=F,
+                          workspace=256e6)
```

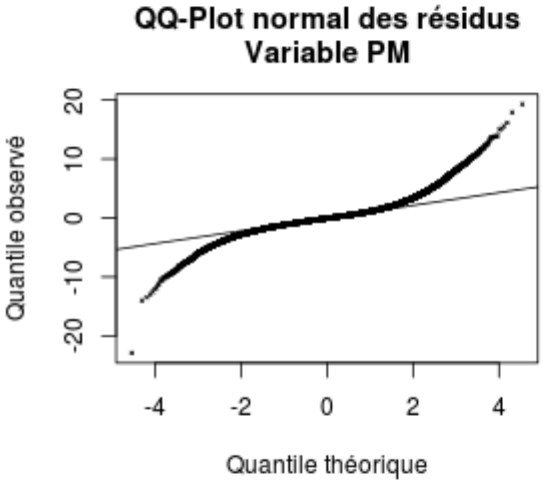
g. Réaliser un modèle BayesB sur géniteurs avec BGLR

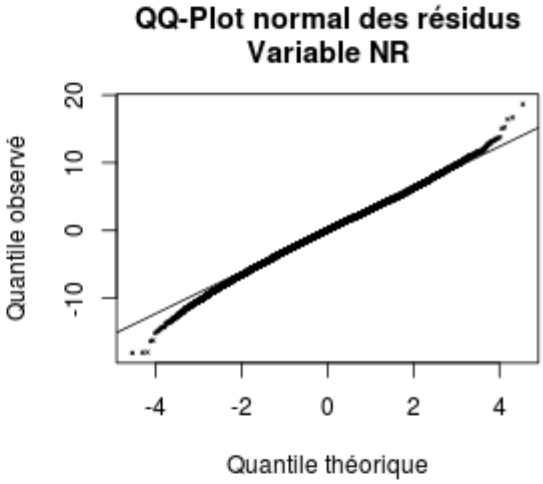
```
> # Paramétrage du modèle Bayésien
> NIT=15000
> BI=5000
> TH=10
>
> # xall contient toutes les données génotypiques, dans l'ordre :
> # jeu d'entraînement – jeu de validation
> xall <- rbind(x_training,x_test)
> ETA<-list(list(X=xall,model='BayesB'))
>
> # y_training.ABW et y_test.ABW contiennent les données phénotypiques
> # de poids moyen (ABW) pour les jeux d'entrainement
> # et de validation respectivement
> y_bglr <- c(y_training.ABW, rep(NA, times=length(y_test.ABW)))
>
> # BayesB univarié sur géniteurs pour le groupe A et la variable ABW
> library(BGLR)
> fmBB.ABW <- BGLR(y=y_bglr,
+                  ETA=ETA,
+                  verbose=F,
+                  nIter=  NIT,
+                  burnIn= BI,
+                  thin=   TH)
```

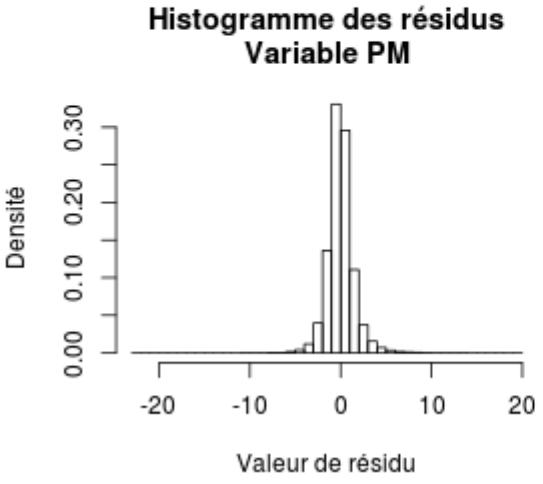
Annexe B : Graphiques pour l'étude des postulats sur les résidus du modèle G-BLUP (G_{EM}) multivarié complet sur descendance

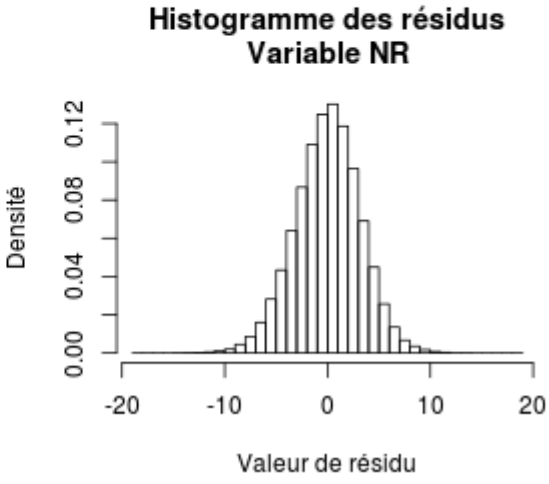




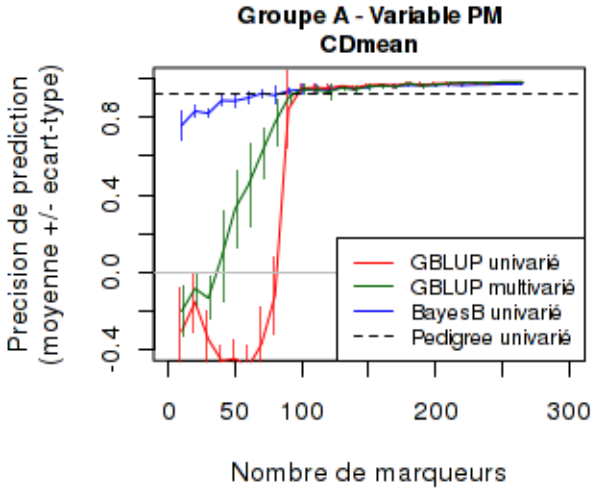


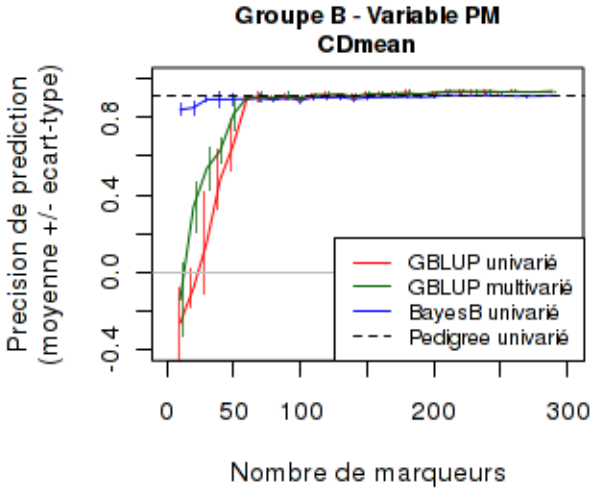


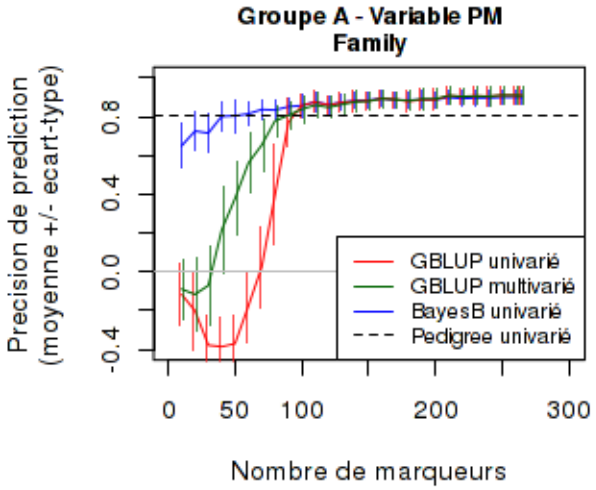


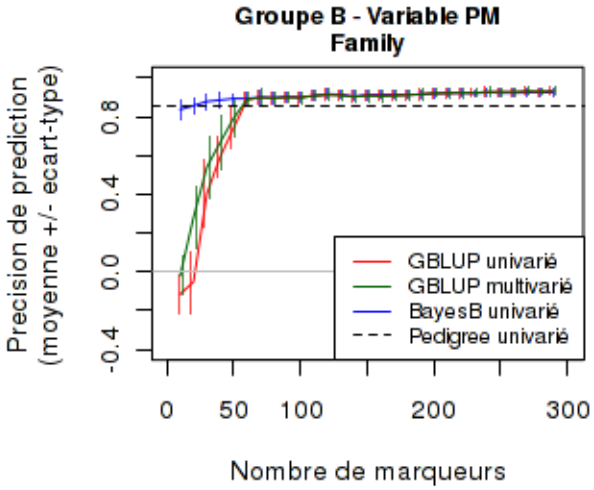


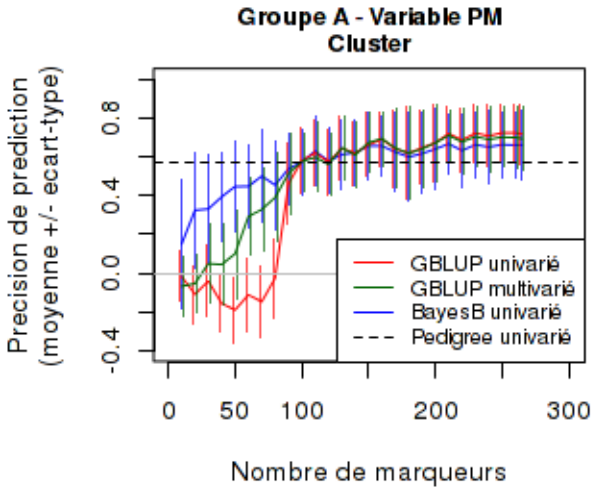
Annexe C : Précision de prédiction des modèles sur géniteurs en fonction de la densité de marqueurs moléculaires, pour chaque combinaison de groupe, de variable et de type de VC

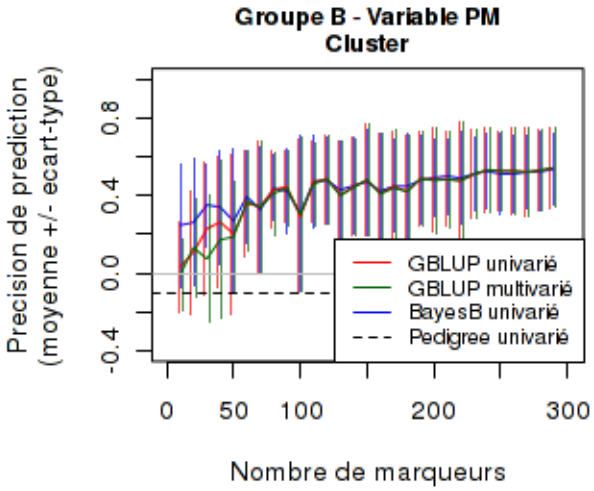


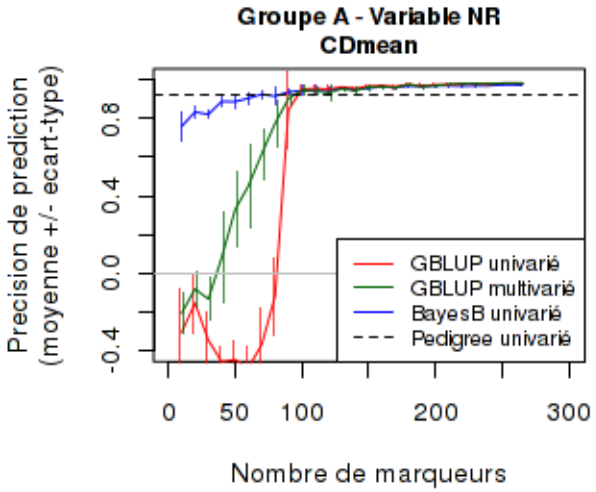


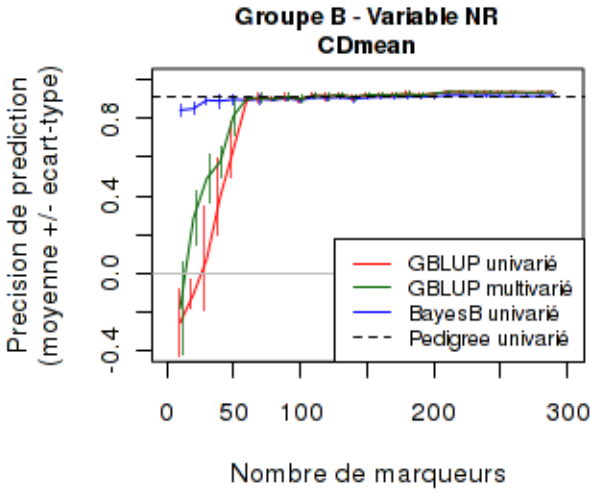


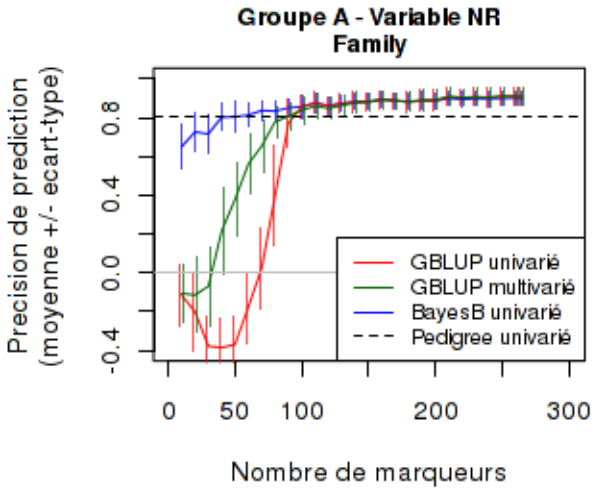


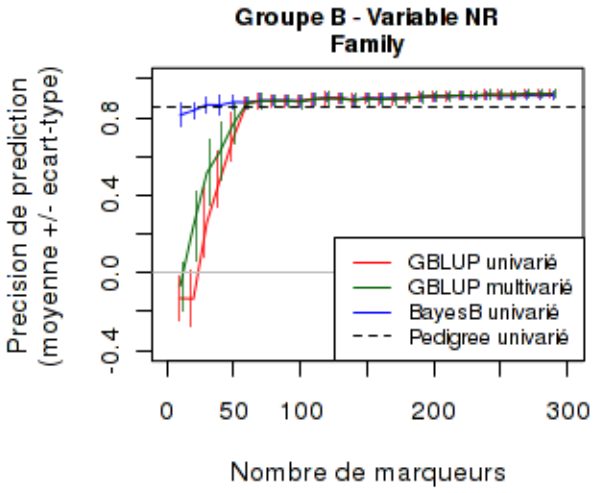


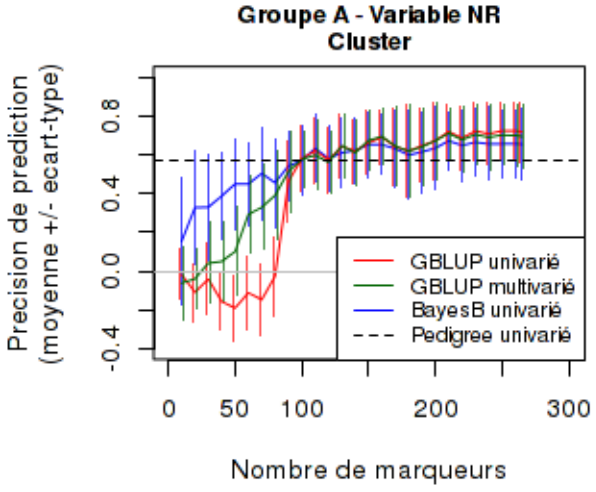


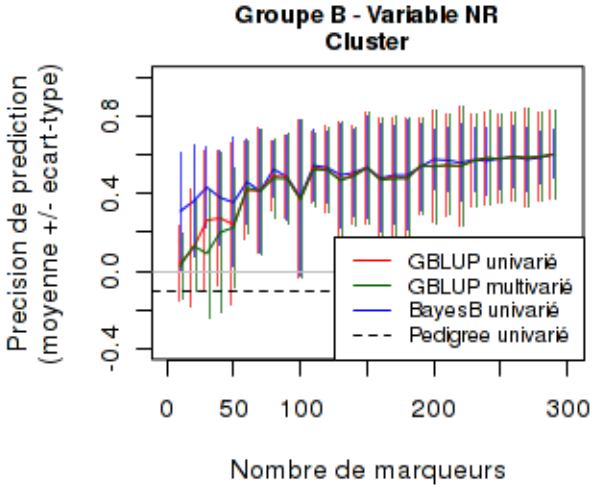

Groupe A - Variable NR
CDmean
Precision de prediction
(moyenne +/- ecart-type)
-0.4
0.0
0.4
0.8
0
50
100
200
300
Nombre de marqueurs
GBLUP univarié
GBLUP multivarié
BayesB univarié
Pedigree univarié


Groupe B - Variable NR
CDmean
Precision de prediction
(moyenne +/- ecart-type)
-0.4
0.0
0.4
0.8
0
50
100
200
300
Nombre de marqueurs
GBLUP univarié
GBLUP multivarié
BayesB univarié
Pedigree univarié


Groupe A - Variable NR
Family
Precision de prediction
(moyenne +/- ecart-type)
-0.4
0.0
0.4
0.8
0
50
100
200
300
Nombre de marqueurs
GBLUP univarié
GBLUP multivarié
BayesB univarié
Pedigree univarié


Groupe B - Variable NR
Family
Precision de prediction
(moyenne +/- ecart-type)
-0.4
0.0
0.4
0.8
0
50
100
200
300
Nombre de marqueurs
GBLUP univarié
GBLUP multivarié
BayesB univarié
Pedigree univarié


Groupe A - Variable NR
Cluster
Precision de prediction
(moyenne +/- ecart-type)
-0.4
0.0
0.4
0.8
0
50
100
200
300
Nombre de marqueurs
GBLUP univarié
GBLUP multivarié
BayesB univarié
Pedigree univarié


Groupe B - Variable NR
Cluster
Precision de prediction
(moyenne +/- ecart-type)
-0.4
0.0
0.4
0.8
0
50
100
200
300
Nombre de marqueurs
GBLUP univarié
GBLUP multivarié
BayesB univarié
Pedigree univarié